Research Report 1332

IMPROVING THE SELECTION, CLASSIFICATION, AND UTILIZATION OF ARMY ENLISTED PERSONNEL

PROJECT A: RESEARCH PLAN

Human Resources Research Organization,
American Institutes for Research,
Personnel Decisions Research Institute,
and
Army Research Institute

SELECTION AND CLASSIFICATION TECHNICAL AREA

U.S. Army
Research Institute for the Behavioral and Social Sciences

May 1983

U. S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

L. NEALE COSBY
Colonel, IN
Commander

Technical review by:

NEWELL K. EATON
JOYCE L. SHIELDS

Notices

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-TST, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Research Project 1332	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Improving the Selection, Classification and Utilization of Army Enlisted Personnel Project A: Research Plan		5. TYPE OF REPORT & PERIOD COVERED Research Report Oct82-Sep89
		6. PERFORMING ORG. REPORT NUMBER RP-PRD-83-7
7. AUTHOR(s) Human Resources Research Organization American Institutes for Research Personnel Decisions Research Institute Army Research Institute		8. CONTRACT OR GRANT NUMBER(s) MDA903-82-C-0531
9. PERFORMING ORGANIZATION NAME AND ADDRESS Human Resources Research Organization 300 N. Washington Street Alexandria, VA 22314		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2Q263731A792
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Avenue Alexandria, VA 22333		12. REPORT DATE May 1983
		13. NUMBER OF PAGES 456
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

The Army Research Institute technical point of contact is Dr. Newell K. Eaton. His telephone number is (202) 274-8275.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Validation, Predictors, Longitudinal Data Base, Validity Generalization, Construct Validation, Army-wide Measures, Job Knowledge Tests, Predictor Measures, Criterion Measures, Performance Measures.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This research report describes a research plan for a project whose objectives are to: (1) validate current and future ASVAB against soldier performance; (2) develop new selection and classification procedures and measures to optimize the match between soldier abilities and MOS requirements; and (3) develop computer-based decision-aids for managers of the Army's manpower processes.

The objectives of the research will be met by: (1) developing new ways to measure and collect data on the military applicant pool; (2) developing and

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. evaluating new predictors of soldier performance (e.g., psychomotor, perceptual, cognitive abilities, and biographical information); (3) developing new methods to measure and analyze training performance; and (4) developing and refining adequate, efficient soldier performance measures and predictors of enlisted personnel and NCO success.

Research Report 1332

IMPROVING THE SELECTION, CLASSIFICATION, AND UTILIZATION OF ARMY ENLISTED PERSONNEL

PROJECT A: RESEARCH PLAN

Accession For	
NTIS GRA&I	X
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A	

Human Resources Research Organization,
American Institutes for Research,
Personnel Decisions Research Institute,
and
Army Research Institute

Submitted by:

Newell K. Eaton, Chief
SELECTION AND CLASSIFICATION
TECHNICAL AREA

Approved as technically adequate
and submitted for publication by:

Joyce L. Shields, Director
MANPOWER AND PERSONNEL
RESEARCH LABORATORY

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333

Office, Deputy Chief of Staff for Personnel
Department of the Army

May 1983

Army Project Number
2Q263731A792

Manpower and Personnel

FOREWORD

This document describes a path toward achieving the goals of the Army's current, large-scale manpower and personnel research effort for improving the selection, classification, and utilization of Army enlisted personnel. The thrust for the project came from the practical, professional, and legal need to validate the Armed Services Vocational Aptitude Battery (ASVAB--the current US military selection/classification test battery) and other selection variables as predictors of training and performance. The portion of the effort described herein is devoted to the development and validation of Army Selection and Classification Measures, and referred to as "Project A". This work is funded primarily by Army Project Number 2Q263731A792. Another part of the effort is the development of a prototype Computerized Personnel Allocation System, referred to as "Project B". Together, these Army Research Institute research efforts, with their in-house and contract components, comprise a landmark program to develop a state-of-the-art empirically validated personnel selection, classification, and allocation system.

EDGAR M. JOHNSON
Technical Director, ARI and
Chief Psychologist, U. S. Army

PREFACE

The planning for this research was initiated by the US Army Research Institute for the Behavioral and Social Sciences (ARI) in 1980. As in-house resources were evaluated, it became apparent that the massive scope of the effort would be best met by a combination of the talents of research scientists and managers from ARI as well as contract research organizations. In 1981 ARI in-house scientists set to work developing the basic research requirements for the effort, and specified those for contract research in a Statement of Work for "Projects A and B". These requirements and specifications were developed by many scientists at ARI, including Joyce Shields, Larry Hanser, Frances Grafton, Hilda Wing, Joseph Zeidner, Newell Kent Eaton, Neil Dumas, and John Mellinger. Prior to project resourcing and contract award on September 30, 1982, the research program was coordinated extensively with the Departments of Defense, Air Force, and Navy, and with the academic and scientific community. Three papers were developed to present the research needs, concepts, and strategies. These were prepared for the Defense Department, the Joint Services, and the Defense Advisory Committee on Military Personnel Testing (Eaton, Wing, Hanser, Dumas, and Shields, November, 1981), the U.S. Army Policy Council (Shields, February, 1982), the American Psychological Association 1982 Annual Meeting (Eaton and Shields, August, 1982).

The primary goal of the project is the empirical demonstration of the relationship of predictor tests to training and performance. To achieve this goal, the most efficient and least disruptive procedure for assessing soldier performance will be developed. Performance measures for soldiers

across all occupations, as well as occupation-specific measures, will be developed and tested. A longitudinal research data base will be created, and contain biographic data, aptitude/achievement indices, training and soldier performance measures, and personnel ratings and actions for a large sample of soldiers over a number of years. This data base will provide the foundation for a system of empirically based quantitative relationships, predicting future performance from past performance weighted according to actual predictive ability for a given time in the career of an individual. For example, initial entry biographical and aptitude/achievement data may be useful for early training decisions, while for mid-career decisions training and initial job performance will be weighted more heavily. Actual weights will be determined empirically.

A second, more ambitious goal is to optimize the match between applicants and occupations. This effort incorporates Army priorities, supply forecasts, and applicants' aptitudes, preferences, and predicted performance capabilities. This goal will be accomplished by mating soldier performance data with an allocation system developed by operations research/computer science professionals using linear and goal programming techniques. Individual and group data provided in the longitudinal data base will be evaluated along with extensive projections of the characteristics of available personnel resources and the Army's personnel requirements based on the types, numbers, and variety of Army occupations. From this, an allocation system will be developed to make the best match of individuals to occupations on a near-real time basis. Constraints include optimizing the system from the Army's perspective, that is, filling critical occupations first, making best use of individuals with unique skills or abilities, and controlling costs.

A consortium, led by the Human Resources Research Organization (HumRRO) and including the American Institutes for Research (AIR), and the Personnel Decisions Research Institute (PDRI) was selected by ARI as the contract organization offering the most innovative and creative approaches to meet the objectives of this project. This research plan represents the contributions and efforts of a great many people, including the ARI planners, Joint Service project evaluators, ARI research scientists and project monitors, members of the Scientific Advisory Group, Interservice Advisory Group, Army General Office Advisory Group, and the principal staff of the consortium of contract research organizations.

Deserving of specific mention for continued stimulation, support, and guidance, are Lieutenant General Maxwell R. Thurman and Major General H. Norman Schwarzkopf.

Members of the numerous groups who contributed their efforts in developing this project were:

Joint Service Project Evaluators:

Dr. Hilda Wing, Dr. Lawrence Hanser, Dr. Bruce Gould, Dr. Martin Wiskoff, Dr. John Mellinger, Dr. Paul Rossmeissl, and Dr. Newell Kent Eaton.

Governance Advisory Group Chairman:

Major General H. Norman Schwarzkopf

Interservice Advisory Group members:

Dr. G. Thomas Sicilia, Dr. Joyce L. Shields, Dr. Martin Wiskoff, and Lieutenant Colonel J. P. Amor.

Scientific Advisory Group members:

Dr. Phillip Bobko, Dr. Thomas Cook, Dr. Milton Hakel, Dr. Lloyd Humphreys, Dr. Robert Linn, Dr. Mary Tenopyr, and Dr. J. E. Uhlaner.

Army General Officer Advisory Group members:

Major General William O'Leksy, Major General Maurice O. Edmonds Brigadier General Gary E. Luck, and Brigadier General John W. Foss.

The staff of the Army Research Institute and HumRRO Consortium responsible for writing, monitoring, reviewing, editing, and approving the research plan, and for the execution of the research:

Army Research Institute		HumRRO Consortium
Newell K. Eaton Joyce L. Shields		Marvin H. Goer John P. Campbell Robert Sadacca James H. Harris
Paul G. Rossmeissl	Task 1	Donald H. McLaughlin Lauress L. Wise Ming-Mei Wang
Hilda Wing	Task 2	Norman G. Peterson
Rebecca L. Oxford-Carpenter	Task 3	Robert Vineberg John Joyner
Lawrence M. Hanser	Task 4	Joseph A. Olmstead Walter C. Borman Barry Riegelhaupt
Michael G. Rumsey	Task 5	William C. Osborn George Wheaton

After months of negotiations, coordination meetings, staff meetings, advisory group meetings, and reviews, this research plan now represents a roadmap leading toward the accomplishment of the project goals. The recommendations of all planners, evaluators, and advisors were considered, and most were incorporated into the plan through the diligent and creative efforts of the contractor staff. Compromises were legion, based on availability of resources, personnel, soldiers to participate in the research, travel, requisite development of scientific underpinnings, and legitimate differences of opinion. Such compromises were negotiated and developed while attending to the priorities, intentions and needs of all concerned.

Troop support required is another example of compromise. The benefits of the project were carefully weighed against the costs of the troop time required. Four tasks required data collection: Predictor measures, training criteria, Army-wide criteria, and MOS-specific criteria development. For each of these, 4-5 sets of data collection efforts involving troop support were originally recommended. To more effectively utilize troop support, and in part to strengthen the research design, some data collection efforts have been merged across tasks in this plan. Troop support locations cited in the plan are tentative recommendations to be coordinated with appropriate organizations.

The plan is also a compromise with time. No research plan is ever complete or unchanging; this one is no exception. It is a snapshot, representing the best picture of the project from the perspective of Spring, 1983. It is intended to be changed, updated, and improved over the years of the project. Semi-annual meetings of the three advisory groups will yield

changing insights, strategies, and needs which will make the project more responsive to pressing operational Army requirements and scientific issues. Collected data will provide insight as to which pathways are proving fruitful and which should be modified or terminated. It is our desire that the project continue to evolve over the years through continued healthy discourse among the Army's senior leadership, representatives of the DOD and Joint Services, the scientific community, and the ARI and contractor scientists. Our aims are: to provide the Army with a greatly improved, empirically based personnel system responsive to the needs of the service, while considering the unique abilities, interests, and desires of individual soldiers; to complete this major project using the best techniques in applied personnel selection and classification research; and, to substantially enhance scientific knowledge in the area.

Newell Kent Eaton

NEWELL KENT EATON
ARI Principal Scientist and COR

Reference Notes:

Eaton, N.K. & Shields, J.L. U.S. Army soldier selection, classification and utilization research program. Paper presented at the annual meeting of the American Psychological Association, Washington, August 1982.

Eaton, N.K., Wing, H., Hanser, L.M., Dumas, N.S., & Shields, J.S. Improving the selection, classification, and utilization of Army enlisted personnel. Paper presented at Department of Defense Advisory Committee on Military Personnel Testing Meeting, San Antonio, November 1981.

Shields, J.L. The Army's personnel system. Paper presented at the Army Policy Council Meeting, The Pentagon, February 1982.

RESEARCH PLAN

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

INTRODUCTION

The overall purpose of the research projects for Improving the Selection, Classification, and Utilization of Army Enlisted Personnel is to enhance the Army's ability to accomplish its peacetime and mobilization missions through improved matching of individuals to military occupational specialties. Toward this goal, the Army Selection and Classification Measures - Development and Validation Project (Project A) is devoted to the development of an expanded and comprehensive selection/classification test battery and the validation of that test battery against a full array of existing and newly developed criteria. Specifcally, Project A is to:

- validate existing selection measures against both existing and project-developed criteria, the latter to include both Army-wide performance measures based on newly developed rating scales and direct measures of MOS specific task performance;

- develop and validate new and/or improved selection and classification measures;

- validate proximal criteria, such as performance in training, as predictors of later criteria, such as

job performance ratings, so that more informed reassignment and promotion decisions can be made throughout the individual's tour;

- determine the relative utility to the Army of different performance levels across MOS; and

- estimate the relative effectiveness of alternative selection and classification procedures in terms of their validity and utility for making operational selection and classification decisions.

Project A is criterion-driven. Its coherence derives from the fact that all of its substantive tasks focus on a single domain, which we can label "effective performance in the Army." Project A must define (state the dimensions or components of) that domain, measure (develop operational criteria for) that domain, and predict (specify the prior information relevant to) that domain. All of the activities of Tasks 1 through 5 must be driven by, and made comprehensible in terms of, the performance components which constitute the domain of interest. The project must not be viewed and must not be conducted as a set of separable tasks that make "inputs" to one another and that are to be "integrated" somehow. Such a view misses the essential unity of the effort; Project A is one project.

NEEDED IMPROVEMENTS IN THE ARMY SELECTION AND CLASSIFICATION SYSTEM

The current Army personnel system has a number of deficiencies which must be addressed in Project A:

1. Predictors covering the full range of the performance domain or criteria space are lacking. Currently, available measures chiefly focus on cognitive abilities. Relevant non-cognitive measures such as psychomotor/perceptual coordination, vocational interest, and biographical indices need to be developed and their usefulness in predicting aspects of Army-wide and MOS specific performance determined.

2. Measures of job performance are lacking. Current measures of job proficiency (SQT) are designed primarily as diagnostic training tools rather than as indicators of successful job performance.

3. The selection and classification of individuals are based on the relationship of entrance tests to performance in training, not performance on the job.

4. The Army does not have the system of data to make critical personnel decisions throughout a soldier's life-cycle based on individual job performance and the needs and priorities of the Army.

5. Currently, if an applicant chooses a specific training program and meets the minimum aptitude requirements, he or she is placed into that training if an opening exists. This procedure does not take into account where that individual could best serve the needs of the Army or even where that individual could be most successful in the Army.

6. The Army does not have efficient means of e[illegible]essing needs and policies in terms of personnel goals, constraints and trade-offs. A dynamic, adaptive, self-adjusting system that supports Army management decision-making is required.

The reasons for these deficiencies stem chiefly from the dynamics in the labor market, the new requirements produced by emerging weapon systems, and the inevitable lag of an operational system behind the most recent technological advances in testing and personnel decision making.

PROJECT A: MAJOR TASKS

Project A is organized into five major research tasks:

Task 1. Validation

Task 1 has two major components. The first is to maintain the data base and provide the analytic procedures to determine the degree to which performance in Army jobs is predictable from some combination of new or existing measures. The second is to determine whether the existing set of predictors, new predictors, or some combination of new and existing predictors has utility over and above the present system. These two components must be accomplished in light of state-of-the-art technology in personnel selection research.

This task encompasses the integrated analysis of all data generated through research activities in the other tasks. While separate teams will be collecting and analyzing subsets of data in accomplishing their tasks, personnel working on this task will be analyzing combined data files. A Longitudinal Research Data Base shall be developed and maintained as part of Task 1.

Task 2. Prediction of Job Performance

To date, a large proportion of the efforts of the armed services in this area have been concentrated on improving the ASVAB. The ASVAB, and other existing test batteries, are primarily indicators of skills that require cognitive abilities. However, many critical Army tasks appear to require

psychomotor and perceptual skills for their successful performance. It is perhaps in such non-cognitive domains that the greatest potential for adding valid independent dimensions to current classification instruments are to be found.

The research plan includes identifying, reviewing and evaluating instruments and variables which may be used at enlistment for predicting Army success.

A critical aspect to this task is the demonstration of the incremental validity added by new predictors. While it may be necessary to rely on content and construct methods of validity in the development phase of new predictors, the Army requires criterion-related field research as support for generalization and extension on findings. Prior to acceptance and use of any new predictors there will be a clear demonstration that they add validity beyond that provided by current predictors and that the cost of test administration, scoring, etc., is justified by the value of the information provided.

<u>Task 3. Measurement of School/Training Success</u>

The objective of Task 3 is to derive school and training performance indices that can be used as 1) criteria against which to validate the initial predictors, and 2) predictors of later job performance. Insofar as possible, these measures will provide information regarding relative standing of trainees both within and across training programs.

The general scope of this task is to evaluate currently available measures and if necessary, to revise them or develop new measures. Comprehensive job knowledge tests will be developed for the sample of MOS investigated and their content and construct validity will be determined.

Task 4. Assessment of Army-wide Performance

In contrast to performance measures which may be developed for a specific Army MOS, Task 4 will develop measures that can be used across all MOS (i.e., Army-wide). That is, the intent is to develop measures of first and second tour job performance against which all Army enlisted personnel may be measured. What is being measured might be termed "soldiering." In fact, a major objective for Task 4 is to develop a model of soldier effectiveness that specifies the major dimensions of an individual's contribution to the Army as an organization. Another important objective of Task 4 is to develop measures of utility that can be used to scale various performance levels across different MOS.

Task 5. Develop MOS Specific Performance Measures

The focus of Task 5 is the development of reliable and valid measures of specific job task performance for 9 selected MOS (out of a sample of 19 MOS). This task may be thought of as consisting of three major components: job analysis, construction of job performance measures, and construct validation of the new measures.

While only 19 MOS will be analyzed during this project, the Army may in the future wish to develop job performance measures for a larger number of MOS.

For this reason, the methods are intended to be generalizable to all Army MOS. Also, the Army must be able, using its existing personnel and resources, to carry out the developed techniques on a regular recurring basis. Finally, the analyses must provide for the establishment of performance standards at both minimum and higher levels.

GENERAL COMPONENTS OF THE RESEARCH PLAN

The specific issues and procedural steps comprising this research effort are detailed in the sections corresponding to Tasks 1-5. However, the full project is organized around a number of major data collection efforts that provide information for several of the tasks simultaneously. They will be large and expensive, but they are fundamental to the success of the project. Without them the Army cannot realize its goals for developing a comprehensive selection and classification system that is statistically, psychometrically, and operationally sound. Consequently, before proceeding to the more detailed research plans, these major procedural components will be described.

Sampling Considerations

There are two sampling considerations. First, we shall select a sample of MOS from the universe of possible MOS; then we shall obtain samples of enlisted personnel (EP) within each MOS. The MOS are the primary sampling units. Large and representative EP samples are important mainly to the extent that they enhance the stability of the statistical results obtained for the sample MOS.

There is a trade-off in the allocation of project resources between the number of MOS researched and the number of subjects tracked within each MOS: the more MOS investigated, the fewer subjects per MOS can be tested and vice versa.

We propose to collect new data for a sample of 19 MOS. To samples from all 19, we will administer the new predictors (from Task 2) and collect the school and Army-wide performance data (of Tasks 3 and 4). To 9 of these MOS, we will also administer the MOS-specific performance measures developed in Task 5. The 9 MOS will be chosen to provide maximum coverage of the total array of knowledge, ability, and skill requirements of Army jobs, given certain statistical constraints.

MOS Selection

Unfortunately, painstaking examination since the start of the contract by Task 5 personnel has not provided sufficient data to permit a confident judgment that any particular sample of MOS is representative of the population of MOS for which personnel decisions must be made. To support such a judgment, one would need job analysis information on the similarities in job requirements and/or job tasks across MOS such that MOS could be clustered into maximally homogeneous subgroups. Since such data do not exist in any systematic form, a first sample of MOS has been drawn by using the following considerations:

1) High density MOS that would provide sufficient sample sizes for statistically reliable estimates of new predictor validity and differential validity across racial and gender groups.

2) Representative coverage of the aptitude areas measured by the ASVAB area composites.

3) High priority MOS (as rated by the Army in the event of a national emergency).

4) Representation of the Army's designated Career Management Fields (CMF).

5) Adequate representation of the types of jobs required to accomplish the Army's mission.

A much more complete specification of the procedure that was used is given in the description of Task 5. However, a summary list of the MOS that were selected is given in Table 1. Summary characteristics of the proposed sample are given in Table 2.

From this list, 4 MOS were identified by ARI and Project A staff and the Project A Governance Advisory Group as encompassing a wide range of job characteristics and as being unlikely to be eliminated as the result of gathering further data. These MOS are indicated in Table 1 by an asterick. They constitute the MOS from which the FY83/84 longitudinal sample will be drawn. Work has already begun on the development of performance measures for these MOS.

Again, within the limits of currently available information this array of MOS represents, as best it can, a) the full population of MOS for which new classification measures would be used, b) the range of aptitudes currently used to make selection decisions, c) high priority MOS, d) MOS that are projected to increase in density, and e) MOS that contain enough people to permit stable estimates of alternative prediction equations and differential validities across racial and gender groups and across MOS.

Without sufficient precision in the statistical estimates all other questions cannot be answered. It is particularly crucial that questions of racial and gender fairness be thoroughly explored at the outset.

Table 1
MOS PROPOSED FOR PROJECT A

MOS	Title	CMF	APT Comp	Priority MOS	Total	FY81 Accessions			Trainee Projections		Expected Number Graduates[1]
						Women	Blacks	Hispanic	FY83	FY84	
• 05C	Radio TT Operator	31	SC	No	3175	585	898	119	2004	2200	1845
• 63B	Vehicle & Generator Mech	63	MM	No	4653	386	1178	242	5304	4402	4280
64C	Motor Transport Operator	64	OF	Yes	5440	774	1279	141	3706	5000	4484
• 71L	Admin Specialist	71	CL	No	4484	2744	1987	215	6191	4592	3858
13B	Cannon Crewman	13	FA	Yes	5783	0	2053	387	8082	3353	3572
91B	Medical Care Specialist	91	ST	Yes	3074	924	876	224	3781	unav	3621
•19E/K	Tank Crewman	19	CO	Yes	3233	0	604	188	3223	3281	2912
95B	Military Police	95	ST	Yes	6073	704	624	127	5720	5300	4373
11B	Infantryman	11	CO	Yes	7028	0	1128	387	12633	13710	11338
76Y	Unit Supply Specialist	76	CL	No	4565	1179	1998	283	6636	4091	3829
94B	Food Service Specialist	94	OF	No	3859	715	1416	125	5133	5157	4600
12B	Combat Engineer	12	CO	Yes	3707	0	716	147	844	2540	1845
16S	MANPADS Crewman	16	OF	Yes	691	0	206	27	797	1015	815
55B	Ammunition Specialist	55	GM	No	662	171	283	42	620	810	762
76W	Petroleum Supply Spec	92	CL	NO	849	259	559	43	1373	1350	1234
54E	Chemical Operations Spec	54	ST	Yes	557	89	185	41	1012	1247	1068
67N	Utility Helicopter Rpr	67	MM	No	1032	33	68	29	572	465	470
51B	Carpentry/Masonry Spc	51	GM	No	602	6	136	14	120	483	341
27E	Tow/Dragon Rpr	27	EL	No	33	40	76	17	312	308	258
	Total				68591	8609	18001	2758	68052	69511	56306

[1] Weighted average of Trainee Projections (3 months of FY83 and 9 months of FY84) adjusted for expected school attrition (actual FY81 rates).

Table 2

CHARACTERISTICS OF THE PROPOSED MOS SAMPLE

FY81 ACCESSIONS		PERCENT OF TOTAL	MOS SAMPLE	PERCENT OF TOTAL
TOTAL	133,192	–	58,591*	–
FEMALE	19,757	14.8	8,609	14.7
BLACK	36,034	27.0	16,001	27.3
HISPANIC	6,416	4.8	2,758	4.7

*SAMPLE = 44% OF TOTAL ACCESSIONS

REGULAR ARMY ONLY

Consequently, the selection of the first sample of MOS began with a consideration of the number of people in each MOS for a particular cohort and then tried to maximize coverage of CMF and aptitude areas.

However, besides being statistically reliable, the estimates of selection and classification equations based on data from the 19 MOS must also be evaluated in terms of how appropriately they can be used to make selection and classification decisions for MOS not among the 19 to be researched. This is the classic problem of validity generalization. That is, given empirical validation data for some specific set of jobs (MOS in our case), to what extent can these data be generalized to estimate the validity of the selection measures for jobs (MOS) that have not been analyzed? It is a fact that there are over 250 enlisted MOS in the Army for which selection and classification decisions must be made. It is also a fact that Project A can empirically validate new selection and classification measures in only a small subset of the total number of MOS. There is no perfect way to select the perfect set of MOS so as to precisely maximize the degree of validity generalization. The problem must be approached by multiple methods over the course of the Project. The methods to be used will be described below, after the general nature of the data collection has been described.

The FY81/82 Cohort

In addition to collecting data from new samples, the project will make use of existing file data that have been, or can be, accumulated for 1981 and 1982 accessions. The editing and merging of data from the accessions and

EMF files for entry into the Longitudinal Research Data Base (LRDB) is already well along and will be ready for analyses beginning in late March or early April, 1983. The overall objective is to accumulate as much data as possible on available predictors and available criteria. Henceforth this source of data will be known as the FY81/82 cohort.

There are several factors that argue for an extensive analysis of the available file data for the FY81/82 cohort:

- o These are the best data currently available for evaluating the validity of the current form of the ASVAB (8, 9, 10). Therefore, there are a number of basic validation questions for which the EMF and accessions file should be useful (e.g., How does the validity of the existing area aptitude scores compare to alternative composites derived from the ASVAB subtests?).

- o If training, EER, SQT, or other archival data are available in sufficient quantity and quality to constitute usable criteria, then the file data can be used as a benchmark against which to compare the incremental validity generated by Project A. That is, for the current predictors and the available criteria, such test validities, composite validities, differential validity across groups (e.g., race) and

different validity across MOS (i.e., validity generalizeability) can be determined. The question is then how much these indices change when the new experimental battery is tried out with the broader range of criteria.

- Analysis of the FY81/82 cohort will allow us to try out a number of new analytic techniques so as to determine if they will be useful in later phases of the project. For example, simultaneous estimation techniques could be used to determine how many significantly different regression equations are needed to predict criterion scores in different MOS. Also, empirical Bayesian techniques could be used to estimate the common regression line across MOS or across cohorts.

Collection of New Data Within MOS

There will be five major new data collections involving three major samples. These furnish much of the information to be used to answer the specific questions posed in the following sections. The sample composition designates subjects by federal fiscal year of entry into the Army. The schedule and types of data collected are for each sample are shown as Figure 1.

Figure 1

Summary of Data Collections and Samples

Data Collection		Sample		
		FY 83/84 (Longitudinal)	FY 83/84 (Concurrent)	FY 86/87 (Longitudinal)
(1)	10/1/83 - 6/30/84	Preliminary Battery		
(2)	6/1/85 - 9/30/85	Trial Battery 1st Tour Criteria	Trial Battery 1st Tour Criteria	
(3)	3/1/86 - 2/28-87			Experimental Battery
(4)	6/1/88 - 9/30/88	2nd Tour Criteria	2nd Tour Criteria	1st Tour Criteria
(5)	2/1/91 - 3/31/91			2nd Tour Criteria

Sample sizes and use of subjects discussed in the following sections and in the individual task descriptions reflect a standardized approach, rather than differentiating sample sizes by MOS in detail every time a data collection is described. This is done in the interest of clarity and economy of discussion. Specific sample requirements, by MOS, by utilization will be detailed in each Troop Support Request and are currently estimated for the FY83/84 cohort in Table 3 in this Introduction.

A schematic of the data collection plan is shown as Figure 2.

Data Collection 1

This first major data collection follows a longitudinal design. New recruits will be tested with a preliminary predictor battery, developed in Task 2, beginning in the late summer or early fall of 1983 and continuing until the summer of 1984. The recruits will be sampled from 4 MOS (05C, 19E/K, 63B, 71L). The principal criterion data will be training

Figure 2

THE OVERALL DATA COLLECTION PLAN

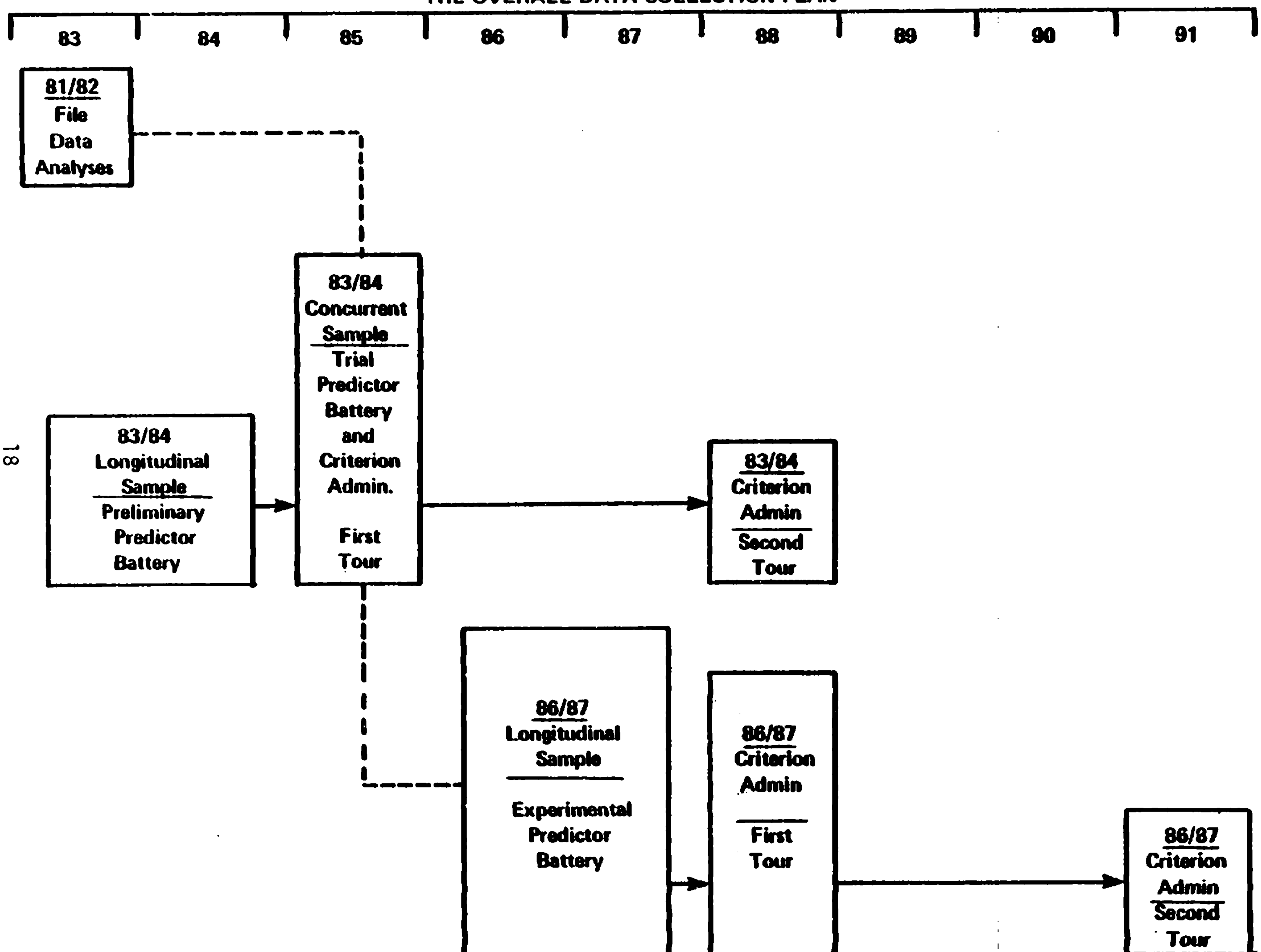


Table 3

ESTIMATED REQUIREMENTS FOR MEASUREMENT OF 83/84 COHORT BY MOS

						Concurrent Validity		Predictive Validity		
MOS	Title	Projected Input	% Complete 1st Tour	% Re-Up	% at Sites	No. Tested at Sites (1st Tour Performance Predictors)	No. Tested at Sites (2nd Tour Performance)	No. Tested at AIT (Predictors)	No. Tested at Sites (1st Tour Performance Predictors)	No. Tested at Sites (2nd Tour Performance)
05C	Radio TT Operator	2150	57	16	37			2158	380	100
63B	Vehicle & Generator Mech	4630	63	08	41			4065	650	100
64C	Motor Transport Operator	4370	64	21	37	826	80			
71L	Admin Specialist	4990	63	26	28			4576	650	100
76Y	United Supply Specialist	4725	62	28	38	735	100			
91B	Medical Care Specialist	3760	65	25	33	645	65			
94B	Food Service Specialist	5150	55	20	39	865	100			
95B	Military Police	5405	70	20	19	675				
11B	Infantryman	13,440	58	17	50	860	100			
12B	Combat Engineer	2115	68	17	42	485				
13B	Cannon Crewman	4190	59	18	52	785	100			
16S	MANPADS Crewman	960								
19E/K	Tank Crewman	3250	61	17	49			2720	650	100
55B	Ammunition Specialist	765	61	26	44	165				
76W	Petroleum Supply Spec	1355	64	23	43	295				
54E	Chemical Operations Spec	1210	59	21	42	240				
67N	Utility Helicopter Rpr	490	76	25	43	130				
51B	Carpentry/Masonry Spec	392	70	10	40	80				
27E	Tow/Dragon Rpr	309	66	24	45	76				

school achievement measures (developed in Task 3) administered as enlistees pass through their training courses. However, the criterion administration sites for the FY83/84 concurrent sample will later be chosen so as to maximize the probability that an individual in the FY83/84 longitudinal sample will fall into the FY83/84 cohort (concurrent sample), first tour sample, which would result in additional criterion measures being available. Since the data collection will constitute a major test of whether previously developed predictors from major domains not covered by ASVAB can add to the prediction of training school grades and other available criteria, a large number of cases will be needed (see Table 3).

Data Collection 2

The collection of data on new predictors, job knowledge tests, and the Army-wide and MOS-specific performance measures will be accomplished in a large field administration of these instruments on the FY83/84 cohort-first tour during 6/85-10/85. The target will be to collect data on the new predictors, job knowledge tests, and Army-wide performance measures for an average of 500 enlisted personnel (EP) in each of the 19 MOS identified earlier; and to collect data on MOS-specific measures for the EP in the 9 MOS of this group for which hands-on instruments will be constructed initially. These data would be used along with the existing preinduction test scores, school grades and behavioral indices already in the cohort data base to validate the ASVAB and other existing measures, conduct a concurrent validation of the new predictors and proximal criteria, improve the psychometric quality of the new instruments, help guide further instrument development, and select the most promising new predictors for administration to the FY86/87 cohort.

Data Collection 3

A longitudinal prediction sample will be collected from the FY86/87 cohort by testing recruits with the revised predictor battery and obtaining school data beginning in March of 1986 and continuing until February 1987. Recruits will be sampled from the 19 focal MOS. (Data may be collected from additional MOS in order to allow better validity generalization from the sample to the population of MOS.) Since this sample will be followed up for purpose of collecting criterion information once during 1988 (first tour) and again during 1991 (second tour) the expected attrition in the sample will be considerable. The expected attrition for a typical MOS is shown in Figure 3. This dictates that it is highly desirable that about 2200 recruits be tested from each MOS on the average. There will most likely not be that many accessions per year for all MOS. In MOS with fewer accessions, we need to obtain as many of the available recruits as possible.

Data Collection 4

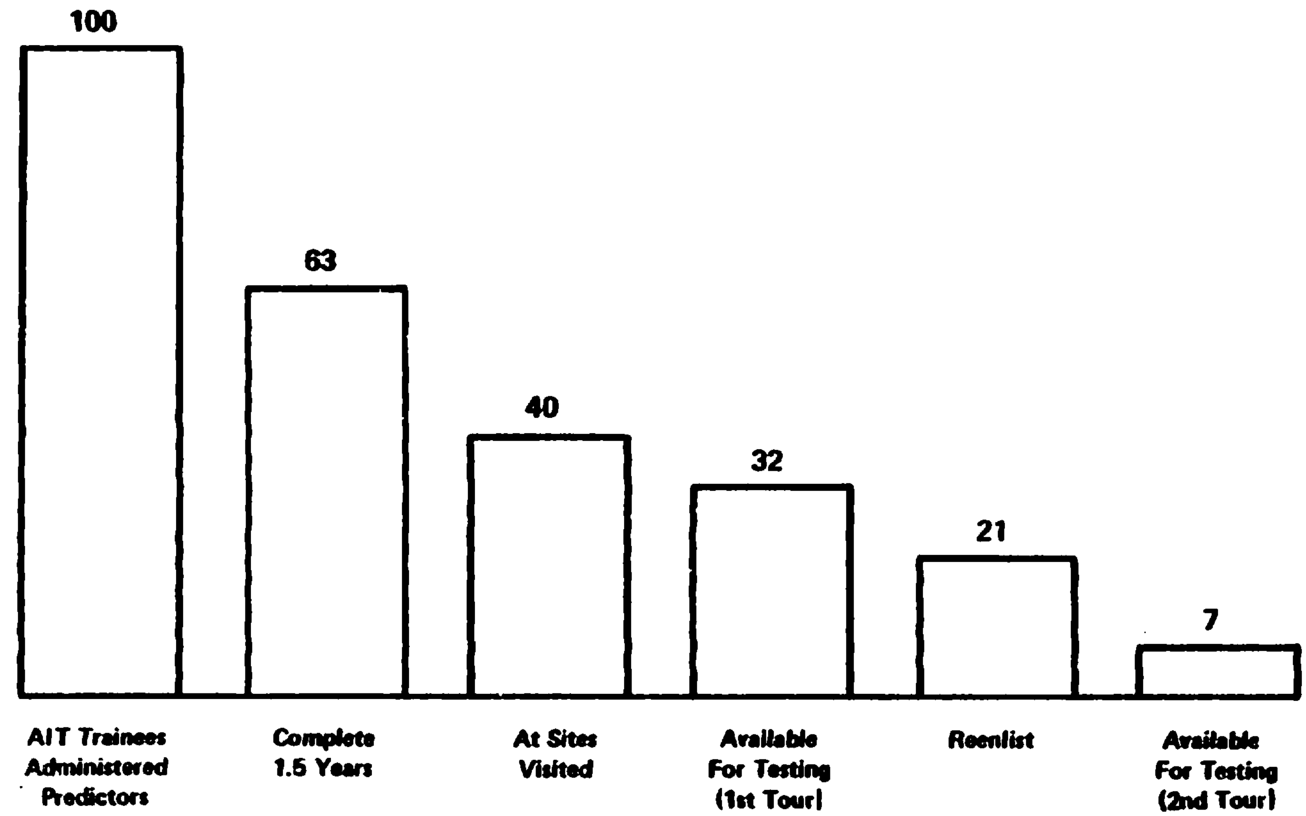
During the period June, 1988 through September 1988, Army-wide and MOS-specific performance measures will be collected at 12 to 15 sites from the FY83/84 cohort which will be in its second tour and the FY86/87 cohort which will be in its first tour.

Data Collection 5

From January 1991 to March, 1991 Army-wide and MOS-specific criterion data will be obtained from the FY86/87 cohort which will be in its second tour.

Figure 3

PERCENT OF TROOPS AVAILABLE FOR PERFORMANCE MEASUREMENT RESEARCH IN A TYPICAL MOS



Sample Size Requirements

The magnitude of the above data collection may seem large. However, it is dictated by the following considerations:

- The overriding goal is to develop a comprehensive selection and classification system that will be implemented across all non-classified enlisted MOS that are associated with advanced instructional training. Consequently, the different parts of the system cannot be studied piecemeal. If the system and connections are not studied as a whole, it will not be possible to develop the optimal set of pre-induction tests, performance measures, and algorithms that link the parts. We must have a large amount of information on each person and this means that sample sizes must be large to insure statistical reliability.

- It is necessary to examine the differences in regressions, correlations, and other statistical indices between gender groups, racial groups, MOS, etc. As has been frequently demonstrated, testing differences between regression and/or correlation coefficients requires very large sample sizes.

- It is necessary, for implementation of the selection and classification system, to draw conclusions about

the level of validity for each MOS. Thus, each MOS that is included must have a sufficient sample size to make reliable statistical conclusions. Since the Army is a large organization, the number of MOS that are researched must be representative of the full range of jobs.

There is considerable attrition from the sample as the cohort moves through its tour. The attrition can be summarized by the following points.

1) A certain percentage of recruits who begin AIT will not finish. Attrition during training is not random, either by MOS or by ability level within MOS.

2) Of those who finish their AIT, a certain percentage will attrite during the first 1-2 years of their tour.

3) Since, for purposes of this project, the criterion assessment of people must take place on a relatively small number of installations, not all the sample will be found on those bases (some will be scattered across a much larger number) and a further reduction in the sample will occur.

4) It is also true that during a given time period, at a given base, not all of the people in the sample will actually be available for testing (e.g., due to leaves, illness, etc.) and additional shrinkage in the sample will occur.

5) Only a small proportion of the original sample will re-enlist and be available for the second tour measures.

6) Of those who re-enlist, only a certain percentage will be on the bases where the testing is taking place at any designated time and be available for testing.

7) The attrition rates over the various stages in a soldier's tour, from AIT to reenlistment are not the same for all MOS. In fact, they vary a great deal, which makes the process of sample selection difficu't.

Estimates on attrition and sample shrinkage for the MOS listed in Table 1 are shown in Table 3 and Figure 2. The estimates are based on actual figures for previous or current accessions. As such, they constitute our best estimate for how these decay functions will look in the future. The initial samples that are required can then be generated by working backwards from the sample sizes that are necessary to provide a minimum level of statistical reliability at the crucial data collection points. The specific sample sizes for each MOS for each major data collection were generated in this way.

In sum, the aim of the project is to develop an organization-wide system for a very important function. In an organization as large and as varied as the Army there is no way that can be done on a small scale.

VALIDITY GENERALIZATION

Before the Computerized Allocation System (CPAS) can become operational, the appropriate parameters in the selection and classification model must be estimated for each MOS in the system. The parameters of interest are the choice of specific tests and the relative weight for each test that will be used to obtain a predicted performance score in a particular MOS for a particular individual recruit. As noted previously, this will involve parameter estimates for over 250 MOS. However, Project A can collect empirical validation data on only 19 MOS. How then can the empirical estimates for 19 MOS be generalized to 250+ MOS?

This is not a problem unique to the Army. It arises anytime that an organization seeks to use a selection test or prediction equation beyond the specific kind of job or situation for which it was validated. Since in any complex organization it is virtually always too expensive and seldom feasible to validate selection measures for every situation in which a decision must be made, the problem arises with considerable frequency.

There is now a growing literature on validity generalization and it is apparent that, for cognitive ability tests at least, validities are much more generalizable across jobs and situations than previously thought. However, there is no simple or universally accepted method by which the parameters of the prediction model can be estimated, a priori, in new situations. It is seldom possible to use classical statistical inference in a straightforward manner. Consequently, to use validity generalization operationally one must somehow use multiple methods to establish the similarity of job

tasks and job requirements across situations such that the validation data (e.g., the multiple R of tests A and B with criterion Y) acquired in one setting (e.g., MOS) can be used to make decisions in other settings judged to be similar.

MOS Clustering

The sample of MOS for the FY83/84 cohort will be based on the considerations previously described. Questions of statistical reliability for estimates of individual test validity and differential validity across racial and gender groups are paramount as well as questions concerning how well the 19 selected MOS represent the larger population of MOS in terms of tasks performed and skills and abilities required.

At the present time there are not available sufficient job analysis data to permit a formalized clustering (via a cluster analysis or factor analysis method) of MOS according to their relative similarity of task content or job requirements. Such a data base will be built before the experimental predictor battery is administered to the FY86/87 cohort. By that time the population of Army MOS will be clustered into homogeneous subgroups such that the array of 19 MOS for the FY83/84 cohort (listed in Table 1) can be evaluated in terms of how well it represents the cluster structure for the population. If there are gaps, the sample of MOS can be adjusted so as to permit as much validity generalization as possible.

The specific steps to be taken to obtain the cluster solution for the FY86/87 cohort will, in part, be a function of the results of a pilot

project now being conducted within Task 5. In that research expert judges are being used to cluster a sample of 111 MOS into subgroups that are homogeneous in terms of their judged similarity in task content. A multi-dimensional scaling algorithm will then be used to recover the dimensions that seem to form the basis of the clustering. Initial analyses of the data being collected contributed to the determination of the list of proposed MOS given in Table 1. The results of this research should give an indication of the feasibility of asking judges to make such judgments, the degree of inter-rater agreement, the number of dimensions it is feasible to use, and the level of detail required for the MOS job descriptions.

If there are no counter indications, then larger panels of experts (of at least two types: personnel professionals and army managers) will be used to cluster all MOS, based on similarity of job content. To do this each MOS will be rated on a standardized set of job content dimensions and job requirement dimensions. It is the profiles of ratings that will be used by the clustering algorithm to generate clusters of MOS that are maximally homogeneous within clusters. The MOS at the centroid of each cluster would be focal MOS. The more similar are the MOS in a cluster the more appropriate it is to use a prediction equation developed on the focal MOS to make selection and classification decisions for all the MOS in the cluster. Also, by varying the weights assigned to the rating dimensions used we can note the degree of similarity between the clusters obtained using job content dimensions vs job requirement dimensions vs using both sets of dimensions. The question as to whether there would be greater validity generalization within MOS clusters formed through job content or

requirement dimensions or through some combination of both types of dimensions can thereby be examined empirically.

FY81/82 File Data Simulation

To a certain degree, the problem can be simulated on the FY81/82 file data. For that data base ASVAB scores and personal history data are available as predictors; and training school grades, SQT scores, and EER ratings (although flawed) are available as criteria on a much larger number of MOS than 19. Thus the prediction equations developed on 19 focal MOS can be applied to each of the other MOS to determine how much information would be lost *if* validity generalization were used rather than an empirically developed prediction equation for the MOS. While such a simulation cannot include psychomotor and other noncognitive predictors because they are not part of the file data, it will portray the validity generalization to be expected for predictors like those already in use.

Simulations Within the Focal MOS

Project A *will* collect complete data for 19 MOS. These 19 can in turn be used to simulate a population of MOS. They then can be clustered into homogeneous subgroups, various methods can be used to estimate empirically a prediction equation for one or more of the MOS in a subgroup, and the information loss incurred by using that equation for the remaining MOS in the subgroup can be calculated.

The effect of subgroup heterogeneity on the amount of information loss can also be explored by systematically increasing the size of the subgroups (and the number of MOS whose data is pooled) and noting the information loss when the resultant equation is applied to the MOS included as well as excluded from the data pool.

Similarity Scaling

Once the tests that will be used in the FY86/87 validation sample are identified it will be possible to carry out another kind of scaling investigation that will address the question of similarity in prediction equations across MOS. Personnel psychologists who understand the ability domains must be used as judges and considerable care must be taken to develop thorough descriptions of each MOS in the array to be scaled.

If the 19 MOS for which we will have extensive predictor and criterion data are considered the "focal" MOS then the relative similarity of each focal MOS to every other MOS can be scaled using psychologists as judges. The MOS should be representative of the cluster structure previously identified. The judges would estimate similarity on the basis of the relative amount of each major ability factor (as determined by analysis of the experimental predictor battery) required by a particular MOS in comparison to each of the 19 focal MOS. Thus there would be a similarity profile (across the major ability factors) that could be used to predict the level of validity and the pattern of predictor weights for each MOS not in the research.

The design permits a number of internal validity checks for the scaling procedure and can even be "validated" for each focal MOS by comparing the predicted results to the actual results for the other 18 MOS for which we will have data. Also, for each MOS not in the research there will be 19 estimates of what the validity of the predictor battery <u>should</u> <u>be</u>. If the scaling were <u>perfect</u> the estimates would converge on the same number. Obviously, the results will not be that precise, but to the extent the 19 estimates converge we can be more confident that the estimated prediction equation for an MOS not among the focal 19 will be a reasonable one.

To carry out this research 30 judges will be required and the scaling task will be extensive in terms of time and effort. Pilot work using 5-10 judges will be carried out to determine the most feasible way to describe MOS and to conduct the scaling sessions so as to minimize the time burden.

Data from the previously described clustering research can also be portrayed in the above fashion. That is, for each focal MOS the correlation of the profile of task dimension ratings for the focal MOS with each other MOS can be examined. Again, for each focal MOS, 18 of these correlations can be compared against the actual result. An important research question is whether the relative similarities portrayed by scaling ability requirements are comparable to those obtained by scaling task dimensions. Also of particular interest is whether MOS drawn from the same clusters tend to have estimated prediction equations that are more similar to one another than equations from MOS drawn from different clusters.

Summary

While most likely none of these individual methods will provide a definitive estimate of the prediction equation for an MOS not in the research sample, when taken together they should provide a reasonable and appropriate estimate. In the process, a number of basic research questions about validity generalization will have been addressed and we will be much better prepared to consider future questions of validity generalization as the MOS structure changes.

GENERAL OUTCOMES

The individual task plans speak to the specific operational and scientific outcomes that will be produced by project A. These reflect a number of basic themes that should be kept in mind.

1) Project A will generate a broader and more complete sample of the predictor space than has ever been used in a selection investigation before. The taxonomy of predictors that is established will stand as a reference point for many years to come.

2) Project A will provide the most thorough attempt ever made to develop standardized tests of actual task performance in skilled jobs. The procedure used will stand as a model to copy.

3) Project A will be by far the most thorough test to date of whether success in training predicts success on the job.

4) Project A will provide a state-of-the-art model for how construct validity can be used to study applied problems in selection and performance assessment. It is our belief that the validation strategy used here anticipates how the validity concept will be reformulated in the forthcoming revision of the Joint Standard for the Use of Psychological Tests.

5) Project A will be the first large selection and classification research effort to incorporate utility in the development of operational decision rules. It is a procedure that will most likely be copied many times.

6) Given the broad range of predictors, criteria, and jobs, project A will be the most comprehensive test ever conducted on questions of differential predictability across gender and racial groups.

7) State-of-the-art answers will be produced about the extent of validity generalizability across jobs, criterion measures, and predictor constructs.

The overall conclusion to be drawn from the above is that although Project A will be time consuming and relatively expensive, the scientific and practical payoffs will exceed the costs many times over.

TASK 1

VALIDATION OF RELATIONSHIPS AMONG PREDICTORS AND PERFORMANCE MEASURES

GENERAL PURPOSE

Within the context of Project A, the primary responsibility of Task 1 is to perform validations of the classification measures. To ensure successful validations, the staff of Task 1 also must work closely with the staff of other tasks, particularly in performing the statistical and psychometric analyses that support the identification and development of new measures. In addition, while the Computerized Personnel Allocation System (CPAS) is still under development, the ASVAB composites that are the primary basis for the current selection and classification procedure have to be updated periodically so that they are maximally effective for the present use. Finally, for the Army's resources to be efficiently used in the implementation of the CPAS, the cost-benefit of alternative selection and classification procedures must be evaluated in terms of the utility of their outcomes. Task 1 will carry out such evaluation in coordination with Project B.

In summary, the purposes of Task 1 within Project A are:

(1) to recommend revisions of Armed Services Vocational Aptitude Battery (ASVAB) composites (for versions 8/9/10 and later for versions 11/12/13) as required in the current selection and classification procedures;

(2) to validate the project-developed classification instruments, and to develop accurate prediction models of the future job performance of enlistees;

(3) to generate appropriate inputs to Project B as required for the implementation of the CPAS;

(4) to evaluate the cost-benefit of alternative classification procedures; and

(5) to provide technical support to the staff of other tasks so as to insure the psychometric qualities of project-developed measures and the data adequacy for the validation of the classification instruments.

It should also be noted that another important purpose of Task 1 is to create and maintain a longitudinal research data base to meet the needs of the present project as well as other ARI projects (e.g., Project B). A comprehensive longitudinal research data base plan has been prepared separately so that the content of the present document is limited to analyses that will be executed using that data base.

BACKGROUND ISSUES AND RATIONALE

Each year more than one hundred thousand new recruits are selected, classified, trained, and assigned to perform the nundreds of jobs required for an effective Army. The system presently employed by the Army for making the initial selection and classification decision has a long history. The development of the primary measure currently used in the system--the ASVAB 8/9/10--can be traced back through earlier forms to the ACB-73, the AQB, the AFQT and ACB, the AGCT, and the original Army Alpha.

In order for an applicant to be qualified for initial enlistment into the Army by the present selection and classification system, he/she must meet a number of eligibility criteria, including age, moral standards, physical standards, and "trainability." The latter determination, the most relevant in the current context, is based upon a combination of two sets of criteria: scores attained on the Armed Services Vocational Aptitude Battey (ASVAB), and educational attainment. The ASVAB is currently administered as an entry test at Military Entrance Processing Stations (MEPS; formerly called AFEES), or at Mobile Examining Teams (MET) sites. It is also administered by MET to high school juniors and seniors; these scores are used for guidance counseling, and are also provided to Army recruiters as a means of identifying mentally-qualified recruitment prospects. In addition to ASVAB,

non-high school graduates are administered a short biographical questionnaire, the Military Applicant Profile (MAP), which has been found to be a useful tool for identifying individuals who are likely to be poor risks in terms of probability of completing Army initial entry training.

For applicants who have not previously taken the ASVAB and whose educational/mental qualifications appear to be marginal based on the Army's trainability standards, a short Enlistment Screening Test may be administered to assess the prospects of passing the ASVAB test. Applicants who appear, upon initial recruiter screening, to have a reasonable prospect of qualifying for service are referred either to a MET site for administration of the ASVAB, or directly to a MEPS. MEPS staff complete all aspects of the screening process, including administration of the mental and physical examination. Based on the information assembled, classification and assignment to a particular training activity are made for those found qualified for enlistment.

About 80 percent of Army enlistees enter the Army under a specific enlistment option that guarantees choice of initial school training, career field assignment, unit assignment, or geographical area. For these applicants, the initial classification and training assignment decision must be made prior to entry into service. This is accomplished at the MEPS by referring applicants who have passed the basic screening criteria (mental, physical, moral) to an Army guidance counselor, whose responsibility is to

match the applicant's qualifications and preferences to the Army's current skill training requirements, and to make "reservations" for training assignments, consistent with the applicant's enlistment option.

The classification and training "reservation" procedure is accomplished by the Recruit Quota System (REQUEST), which was implemented in 1973. REQUEST is a computer-based system to coordinate the information needed to reserve training slots for volunteers. One major limitation is that REQUEST uses simple, <u>minimum</u> qualifications for accessions control. Thus, to the extent that an applicant may minimally qualify for a wide range of courses or specialties, based on aptitude test scores, the initial classification decision is governed by (a) his/her own stated preference (often based upon limited knowledge about the actual job content and working conditions of the various military occupations), (b) the availability of training slots, and (c) priorities/needs of the Army. Numerous procedures for improving the system are under development. These include "MOS Match Module" and the previously mentioned Project B Computerized Personnel Allocation System, as well as other smaller efforts.

This review of the current practice suggests that the present selection and classification procedures could be improved by taking advantage of recent technological advances and developments in decision theory. There is a need for developing a formal decision-making procedure that is aimed at maximizing the overall utility of the classification outcomes to the Army.

However, this decision process must allow for the potentially adverse impacts on recruitment if the enlistee's interests, work values and preferences are not given sufficient consideration. There are clear trade-offs that must be evaluated between the procedures necessary (a) to attract qualified people, and (b) to put them into the right slots.

From this perspective, a classification system must be built upon a thorough understanding of what constitutes effective performance in the Army. In addition, a basis for estimating (predicting) an enlistee's future performance from pre-induction information needs to be established. In order to design a formal classification procedure to improve personnel utilization in the Army, we need to have a predictor battery that is maximally valid and can be administered efficiently at the MEPS. While the Army has a long success in making use of selection and classification tests, the prediction system could most probably be further improved, particularly by adding other noncognitive tests (e.g., psychomotor skills and vocational interests). The improved and/or newly developed predictors need to be validated in terms of the incremental utility they will contribute in addition to the existing predictors. Task 1 is devoted to performing such validations, and at the same time empirically developing accurate prediction models to be employed in the Computerized Personnel Allocation Systems (CPAS).

To further improve the effectiveness of personnel utilization in the Army, the personnel allocation system may consider incorporation of information gathered during training of the soldier's earlier career into the prediction of his/her subsequent performance. On the basis of this enhanced prediction model, a sequential/dynamic decision process can be established to systematically update the assignment of enlisted personnel to jobs that will benefit most from his/her current skills and qualifications. In support of building such a decision model, Task 1 is also aimed at validating the additional post-enlistment predictors (i.e., school/training predictors and in-service predictors) against Army-wide as well as job-specific performance criteria.

SPECIFIC OBJECTIVES

Project A is designed to provide an empirical basis for optimal selection, classification, and utilization of Army enlisted personnel. Optimization will be achieved by allocating personnel in accordance with predictions of their performance in different assignments. Within Project A, Task 1 is designed to evaluate both existing predictors and predictors developed by other Tasks in Project A in terms of the extent to which they meet the goal of the project.

The objectives of Task 1 fall into three categories:

(1) Evaluation of Existing Predictors;

(2) Support for the Development of New/Improved Measures; and

(3) Evaluation of New/Improved Predictors.

In the course of meeting these objectives, Task 1 will be responsible for the development and maintenance of a Longitudinal Research Data Base (LRDB), as described by Wise and Wang (1983).

Each of the three major categories of objectives can be divided into specific objectives that will be carried out over the period of the project. Table 1-1 presents the projected dates for accomplishing each of the specific objectives:

Table 1-1. Projected Dates for Accomplishing the Specific Objectives of Task 1.

Objectives	Projected Dates of Accomplishment
1. Evaluation of Existing Predictors	
1.1. Development of Area Composites for ASVAB-8/9/10	March 1984
1.2. Production of Early Reports on validation issues (validity generalization, cultural fairness, and cross-validation)	November 1983 to April 1985 (separate reports)
1.3. Comparison of Computer Administered and paper-and-pencil version of ASVAB	December 1986
1.4. Development (refinement) of ASVAB area composites for forms 11/12/13	May 1987
2. Support for Development of New/Improved Measures	
2.1. Identification of predictor dimensions requiring improvement	August 1983 (input to Task 2)
2.2. Evaluation of existing performance measures (training and Army-wide)	December 1985
2.3. Assistance in design and analysis, and review of reports	throughout the period of the project, as requested (see Research plans for Tasks 2 through 5 for schedules)
2.4. Validation of proposed new and improved predictors (using FY83/84 cohort data)	April 1987
2.5. Support for the development of utility measures of performance	September 1986
3. Evaluation of New/Improved Predictors	
3.1. Generation of input to CPAS	January 1988 to August 1989 (in stages)
3.2. Cost-benefit comparisons of alternative classification procedures	September 1989
3.3. Production of follow-up reports on validation issues (validity generalization, cultural fairness, cross-validation, stability of relationships)	September 1989
3.4. Development of procedures for updating CPAS parameter estimates	September 1989

Evaluation of Existing Predictors will focus on:

(1.1) development of and assessment of the discriminant validity of new area composite scores for ASVAB forms 8, 9, and 10;

(1.2) production of reports on the validity of assumptions involved in use of the ASVAB for personnel classification, including validity generalization, cultural fairness, and cross-validation of predictive relationships;

(1.3) development of and assessment of the discriminant validity of new Area Composite Scores for ASVAB forms 11, 12, and 13.

The initial work in this phase of the project will make use of data on the FY81/82 cohorts of Army enlisted personnel, including a special training data file developed by the Army Research Institute for use by this project. Subsequent analyses will make use of data from the FY83/84 cohorts. A major step will be taken between the first and second area composite score development efforts to include differential utility of performance as the criterion for optimality.

Support for the Development of New/Improved Measures will focus on:

(2.1) identification of areas in which improvements in existing predictors for classification decisions are most needed;

(2.2) evaluation of current training outcome measures and general performance indicators (e.g., EER, discipline actions) as additional predictors of subsequent performances;

(2.3) response to requests for sampling, design, and analytical assistance and review from other Task Leaders;

(2.4) validation of proposed new and improved predictors, employing data on FY83/84 cohort; and

(2.5) analyses in support of the development of measures (in Task 4) of the utility to the Army of several performance levels in different MOS.

The first two of these objectives will aim to make best use of existing data bases to provide information to the project staff charged with the development of new and improved measures. In providing assistance to other Task Leaders, Task 1 will endeavor to coordinate the data base and analytic activities of the other Tasks so as to minimize the overlap of efforts while maximizing the exchange of empirical results. The efforts of Task 1 with respect to FY83/84 data will mainly involve analyses that cut across the other tasks.

The Evaluation of New/Improved Predictors will focus on:

(3.1) generation of inputs to the Computerized Personnel Allocation System being developed in Project B;

(3.2) cost-benefit comparisons of alternative measurement and assignment strategies for selection, classification, and utilization of Army enlisted personnel at various points in their careers;

(3.3) production of reports on the validity of assumptions involved in use of new and improved predictors for personnel classification, including stability of relationships across time, validity generalization, cultural fairness, and cross-validation of predictive relationships; and

(3.4) development and validation of procedures for updating CPAS parameter estimates on the basis of alternative data collection strategies.

The work in this final phase of the project will be based primarily on the longitudinal data collection on the FY86/87 cohorts. All new and improved predictors and criteria developed in this project will be administered to these soldiers, so that longitudinal predictive validation can be performed. The evaluations will be coordinated with Project B so that increments in validity can be evaluated in the context of actual supply and demand constraints on Army personnel assignments. The last objective (number 3.4) will provide the flexibility needed for continuous operation of the CPAS in the face of changing supplies of and demands for personnel with particular knowledge, skills, and abilities.

OVERALL SUMMARY OF THE PROCEDURE

The work of Task 1 has been divided into eight functional Subtasks. The numbering of these Subtasks has been designed to facilitate correspondence with the other Tasks. In addition to the analytical subtasks, the Longitudinal Research Data Base (LRDB) has been included as Subtask 1.1, and Management and Coordination have been included as Subtask 1.6. The Subtasks are listed below.

Subtask 1.1:	LRDB Development and Maintenance;
Subtask 1.2:	Support for the Development of New/Improved Pre-Induction Measures (Task 2);
Subtask 1.3:	Support for the Development of New/Improved Training Outcome Measures (Task 3);
Subtask 1.4:	Support for the Development of New/Improved Army-wide Criteria (Task 4);
Subtask 1.5:	Support for the Development of MOS-specific Criteria (Task 5);
Subtask 1.6:	Management and Coordination with Other Tasks and with Project B;
Subtask 1.7:	Validation of Existing Predictors; and
Subtask 1.8:	Validation of New/Improved Predictors.

Time lines for these Subtasks and their interfaces with other Tasks and Subtasks are described in the Integrated Master Plan for Project A. For the

purposes of this Research Plan, we have organized the Subtasks in terms of the specific objectives described in the preceding Section. Thus, we first discuss our plans for the Evaluation of Existing Predictors, which will be carried out as Subtask 1.7. Second, we discuss our plans for Support for the Development of New/Improved Predictors and Criteria, which will be carried out as Subtasks 1.2 through 1.5. Finally, we discuss the longitudinal validation of the combined new and improved system for personnel selection, classification, and utilization, which will be carried out as Subtask 1.8.

We have devoted most attention to the Evaluation of Existing Predictors and to the Evaluation of New/Improved Predictors, because the research plans of the other Tasks describe the needed analyses for the Development of New/Improved Measures in detail. To reiterate them in Task 1 would only create redundancy in the content of the Project A Research Plan.

PROCEDURE

Task 1 in Project A plays a dual role, (a) carrying out validation analyses to provide the foundation for the CPAS, and (b) supporting the research efforts of the rest of the project. In the former category fall the validation of current measures, including the development of ASVAB area composite scores based on the FY81/82 cohort data, and the validation of the new and improved battery that will be developed over the course of this project. In the latter category fall the development of the Longitudinal Research Data Base and the various analyses needed to support the development of new MEPS-level predictors, training outcome measures, Army-wide criteria, and MOS-specific performance measures.

Given this multiple role, we have divided our research plan into three sections:

Section 1. Validating existing predictors for use in selection and classification (Subtask 1.7 in the Integrated Master Plan);

Section 2. Supporting the development of new/improved measures (Subtasks 1.2 through 1.5 in the Integrated Master Plan); and

Section 3. Validating new and improved predictors for use in the CPAS (Subtask 1.8 in the Integrated Master Plan).

Activities in Subtasks 1.2 through 1.5 will proceed continuously throughout the project, while those in Subtask 1.7 will be replaced by Subtask 1.3 when

we turn to FY86/87 cohort analyses. A number of methodological issues surrounding the validation analyses will be examined concurrently with these activities. As noted earlier, Subtask 1.1 is discussed in a separate document--the Longitudinal Research Database Plan.

Section 1: Validation of Existing Predictors for Use in Army Enlisted Personnel Selection and Classification Procedures

The development of new and improved instruments for prediction of performance must be based on a thorough evaluation of the current procedures. A major effort assigned to Task 1 is to perform analyses of existing data to determine the validity of the existing ASVAB battery, supplemented by currently available background data. This subtask is further defined as the development of area composite scores based on the current ASVAB; that is, the identification of those scoring procedures that make best use of the ASVAB for Army personnel selection and classification decisions, in the context of their present use.

Development of ASVAB Area Composite Scores

The initial major analytical effort in this project will, to a great extent, aim to identify the best area composite scores that can be derived from the current ASVAB. The results may either corroborate use of the current composites or they may indicate new or revised ones. This effort is of substantial significance to the Army, and is a major product of the first 18 months of the project. It is also a "rehearsal" for the subsequent

improvements to the Army selection and classification system to be developed in this project. To achieve the objectives of this initial effort, several methodological issues that will affect the ultimate results of the project will be dealt with.

The initial improvements in the selection and classification system will be undertaken entirely within the framework of ~~the~~ existing ASVAB usage and will make extensive use of the large amount of prior work on ASVAB area composite scores, especially the validations carried out by the Center for Naval Analysis (Maier, 1981, 1982; Sims, 1978; Sims & Mifflin, 1978; Sims & Hiatt, 1981). Nevertheless, using the special data base developed for the Army on the FY81/82 cohort by ARI, we expect to be able to provide significant enhancements to the current state of knowledge concerning the proper use of ASVAB scores for training/job assignments.

The current version of ASVAB (Form 8/9/10) was introduced in October, 1980. There are nine composites that are defined largely based on the validation of ASVAB 6/7. Each of the nine composites is used to determine the qualification of an applicant for one of the nine specific MOS groups. In addition, the Army continues to define the AFQT composite for use in the initial screening of applicants. Another composite (G or GT, General Technical) is also defined, but not associated with specific MOS groups. Of these 11 composites, four are very similar to the composites used in other services; they are generally referred to as MAGE (M: Mechnical Maintenance;

A: Administrative/Clerical; G: General or General Technical; and E: Electronics Repair).

The initial validation of the ASVAB 8/9/10 composites as the Army's election and classification predictors was carried out by Maier in 1981 (loc. cit), employing final course grade and job proficiency tests as criteria. He concluded that the composites defined on the basis of ASVAB 6/7 validations are valid predictors of training success as well as job proficiency. However, the data used to conduct the validation were based on ASVAB 6/7 scores. Later in 1982, Maier (loc. cit.) validated these ASVAB composites with scores on forms 8/9/10 for the Marine Corps, using final course grades as the criterion. In general, his results confirm the predictive validity of the composites. Other studies, such as those by Sims and his colleagues (loc. cit.) also substantiate the validity of these composites albeit all based on ASVAB 6/7 data.

This brief review of the past validations of existing ASVAB 8/9/10 composites reveals that a complete validation of the current ASVAB tests is still to be carried out. In response to this need, the Army Research Institute has collected a comprehensive set of training performance data on the FY81/82 recruits who were among the first to take ASVAB 8/9/10 and attended Army schools during CY81. The training graduates of this cohort of recruits are now in their first-tour and many have taken the Skill Qualification Test (SQT). Additionally other general performance records

for them are also available from the Enlistment Master File (EMF). Thus it is now possible for us to conduct a validation of ASVAB 8/9/10 tests as predictors in the selection and classification of the Army's recruits, using not only training performances but also the SQT scores and general performance indicators (e.g., EER's, attrition, disciplinary actions) as criteria.

Our approach will be essentially empirical, emphasizing computation of area composites that have been found to be indicative of successful training outcomes and proficient execution of tasks in the field. The initial effort will focus on the predictive validity (absolute as well as differential) of the ASVAB 8/9/10 subtests and the composites currently in use. On the basis of these evaluations, we will then determine whether the effectiveness of ASVAB in the current Army selection and classification practice could be improved either by modifying the existing composites or by developing new ones. Ideally, whether to contin[illegible]ing the current composites or to adopt new ones should be assessed by [illegible] increase of total performance (effectiveness) in the Army as a res[illegible] of basing selection and classification decisions on the revised compo[illegible]s. However, the validation with the FY81/82 cohort data will not be conducted strictly in this context because utility measures for integrating the job performances in the Army into a single effectiveness scale will not be available until 1985. Therefore our present effort will essentially follow the traditional validation approach. When the utility measures are fully developed by this project, the new ASVAB composites will be revalidated more formally.

The FY81/82 cohort validation of the ASVAB 8/9/10 will begin in May, 1983, and conclude in March, 1984. An interim report will be submitted to ARI for review in October, 1983. Based on comments from ARI, we will finalize our recommendations on the set of composites to be used beginning in October, 1984. Because the ASVAB 11/12/13 forms are scheduled for administration in October this year, the area composites will in effect be applied to the ASVAB 11/12/13 scores. Clearly, the present validation of ASVAB 8/9/10 will also have to address the issues that may arise in defining ASVAB 11/12/13 composites. The subtests of forms 11/12/13 are essentially the same as forms 8/9/10, we therefore do not anticipate special difficulties in adopting the revised composites for the new tests. Employing data on forms 11/12/13 to be collected from the FY83/84 and FY86/87 cohort, we will continue the validation of ASVAB in order to assess the validity of the composites using improved sets of criteria and to revise them as required.

In what follows, we first present the objectives of this validation effort, and then describe the procedures that we will follow to accomplish the objectives.

<u>Objectives</u>. The objective of this subtask is to identify a set of area composite scores (or more generally, ASVAB scoring rules) that:

(1) are feasible to implement;

(2) maximize expected performance, when properly implemented; and

(3) exhibit appropriate stability.

Each of these requirements is exceedingly complex and involves controversies that must be addressed. We plan to address them in such a way that the results of this effort will help lay the groundwork for the remainder of the project.

The question of feasibility of implementation concerns such issues as the type of coefficients that can be used in combining subtest scores, the number of subtests in each composite, the number of different composites, the possibility of adjustments based on subsidiary information (such as the "add 10 points for high school graduation" rule proposed by Sims & Hiatt, 1981), and the use of multiple cutoffs for a single MOS. These issues will be addressed by comparing the predictive validities of alternative sets of composites that vary in these respects.

The current ASVAB Area Composites are computed as sums of the subtest standard scores (each subtest is scaled to have mean 50 and s.d. 10) and then converted to a scale that is comparable across the composites (with mean 100 and s.d. 20). As a starting point, we will define composites that employ unit weighting and include 3 or 4 subtests in each composite. Changes in these traditional practices will be recommended only if they result in a significant increase in validity.

Maximization of total expected performance (or allocation average, Brogden, 1946) requires considerations of the constraints of MOS requirements and supply of applicants, as well as a common utility scale for performances in different MOS. The necessary utility data will not be available for this initial effort, so we must rely on the same assumption that previous efforts have incorporated. The assumption is that all measured increments of performance (expressed in standard scales with common mean and s.d. across MOS), either within MOS or between MOS, are equally valuable. The question of requirement and supply will be examined by simulations, i.e., generating data by computer according to the anticipated supply of recruits, making job assignments to match the quota based on the composites, and then evaluating the expected performance of the outcome. These simulations will be coordinated with Project B staff.

Finally, the question of stability involves an appreciation for the costs of altering enlistment procedures, as well as the statistical sophistication to differentiate between real and chance variations. In order to differentiate between real and chance variations, we plan to conduct careful, and extensive, cross-validations.

Procedures. Our procedures for developing the area composites include several major steps. First, preliminary analyses will be performed to determine the availability and adequacy of data to support the validity analysis (e.g., sample sizes, kinds and characteristics of criterion measures, and score scales). Second, we will address the methodological

issues concerning the problem of selectivity (restriction of range), and the possibility of nonlinear relationships between predictors and criteria. The results of these investigations will be used to make appropriate data adjustments and/or transformations so that proper models can be applied to conduct the validation. In addition, the issue of fairness will be examined to insure that the area composites to be developed are valid for groups of special interest. Third, the number of area composites required to reliably differentiate the performances between MOS will be estimated by clustering the MOS into homogeneous groups such that there will be substantial validity generalization among MOS within group, but differential validity across groups. Fourth, validity analysis will be conducted for each MOS group to define the best area composite for that group, and to explore the impact of different cutoff scores for selection into each MOS. Finally, we will carry out cross-validations to lessen the impact of chance variations on the area composites. We now turn to describe the details for each of these steps.

Step 1: Conduct preliminary analyses.

The most important question we must address before the validation concerns the availability and validity of criterion measures. The salient fact is that although all, or nearly all, enlisted personnel take the ASVAB under controlled conditions at a specified time, there is no similar uniformity of criteria.

That this subtask can even be considered at this time is only possible because of the major effort undertaken by ARI to develop a data base of training outcomes on the FY81/82 cohort. Nevertheless, those training outcomes were recorded at various sites around the country, under various conditions, and it will not be appropriate to treat these data as uniform without further documentation. Work already undertaken by ARI has indicated some limitations of the training data.

The meaning of final course grades that constitute the primary criterion in previous validations has also changed as a result of recent emphasis on objective-based training and mastery testing. Based on what we have learned from the training performance data so far, we anticipate that for many MOS, the criterion scores will not have sufficient variability to support meaningful validity analyses. Moreover, unless we rescale the grades to make them comparable across courses/classes within an MOS, so that data from different classes can be pooled for the analysis, we may not have sufficient sample sizes to obtain reliable results.

Therefore, we have begun to examine the similarity of the performance scores between classes and schools using the information provided in the ARI documents and later to be supplemented with information that Task 3 staff are collecting during their school visits. Once it can be determined that the course contents are similar and the tests used are comparable, we will pool those data for actual analysis. We will also perform descriptive analyses by school and by MOS, when the data editing is completed, in order

to assess the score distribution and determine the sample size available for analysis. The sample size is particularly of concern when we want to do subgroup analysis in the examination of the fairness issue. Additionally, we will standardize the course grades within each MOS so that the expected performances will not reflect potentially large between MOS differences. In this regard the ASVAB may be used as a common referent to "equate" course grade or other criterion distributions across MOS.

We intend to supplement training outcome data with later performance measures, such as EER and SQT scores. Because the FY81/82 cohort will have been in the service for approximately 2.5 years at the time we carry out these analyses, it may be possible to compare the area composites based on training outcomes to area composites based on field performance measures. For the SQT scores, we will perform preliminary analyses to assess the effects on soldiers' performance of time intervals between completion of training and administration of the test. The results will inform us whether we should take into account this time factor in validating the ASVAB tests with SQT scores as the criterion.

Step 2: Address methodological issues and make data adjustments.

(1) Problem of selectivity bias (restriction of range).

The most serious issue that confronts us in conducting the validation in the existing Army setting is the problem of selectivity. Because ASVAB composites were used to select and assign the recruits, we expect restriction of ranges in the subtest scores as

a result of such implicit selection (note that this selectivity problem is further complicated by the fact that the recruits can choose among MOS for which they are qualified and there are guarantee and bonus options). This selectivity problem can distort the relationship between the predictors and the criterion if the selection variables are not all included in the prediction equation, and can even introduce nonlinearity (Heckman, 1979).

The traditional approach to alleviating the effects of range restriction on a validity investigation is to make corrections based on the assumptions of linear regression and homoscedasticity. There are two ways to make the corrections -- univariate model and multivariate model. Sims and Hiatt (1981) conducted a simulation to assess these correction methods and found that the multivariate model is more effective and quite satisfactory in reducing the errors of correlation estimates based on restricted samples. The multivariate correction was formulated by Lawley (1943) and is not difficult to apply. We plan to carry out adjustments of the correlations using the FY81/82 applicant population as the reference (base) population.

For the case of a dichotomous criterion, a classical solution for the explicit selection case is available and is due to Gillman and Goode (1946). Regrettably, there does not seem to be any completely satisfactory solution to the dichotomous criterion for the implicit selection case. Because this is the case in which the ASVAB is validated using a dichotomous criterion, it also requires attention when we use a general performance indicator such as attrition as the criterion for validation.

The existing literature on correcting for restriction of range contains solutions for the relatively simple cases; we need to investigate the applicability of these solutions for the present project. In addition, we will investigate the degree to which the statistical assumptions underlying these correction procedures are violated and develop ways of coping with difficult cases.

There have been a few Monte Carlo studies on the effect of using the classical methods of correcting for selection when the assumptions are violated (Novick & Thayer, 1969; Rydberg, 1963; Meredith, 1958; Srinivasan & Weinstein, 1973; Greener & Osburn, 1980). In general, it appears that using the correction formula is better than working with the uncorrected correlation, but the performance of the formula worsens as the degree of selection increases. For low to moderate selection (fewer than 40 percent rejected) the corrected correlation is considerably better than the uncorrected correlation. However, certain patterns of heterogeneous error variance and curvilinear relationships result in unacceptable overestimates of the population correlation (Greener

& Osburn, 1980). We will use the existing data on the FY81/82 cohort to estimate selection ratios, and evaluate the assumptions of homogeneity of error variance and linear relationships. If one or both of these assumptions are violated, we will search for linearizing transformations.

(2) Assumption of linear relationships.

The second issue we must deal with is the appropriateness of linear models for the validation. In general we assume that linear models will be appropriate or at least a good approximation of other models. As noted earlier, selectivity problems may introduce artificial nonlinearity. It is also known that measurement errors (unreliabilities) of the regressor variables frequently distort the underlying regression by introducing nonlinearity into the model (Cochran, 1970; Lindley, 1947). For these reasons, we need to check on the linearity assumption. When the data suggest nonlinearity, we will attempt to linearize it by transformations or by incorporating polynomial terms into the model. However, it should be noted that empirical research frequently finds that non-linear relationsnips are often quite unstable and cannot be replicated. Further, previous research also suggests that most non-linearities can be satisfactoriy approximated by polynomial functions that in effect render the model additive (linear).

(3) Moderator effects and the issue of cultural fairness.

The third issue that is of great policy interest is the question of cultural fairness of the selection and classification procedures. For a successful validation, the predictor and criterion measures should be reliable and free from socio-cultural bias (ideally, from bias against any individual).

In accordance with the current law, fairness or unfairness can be defined as "when members of one race, sex, or ethnic group characteristically obtain lower scores on a selection procedure than members of another group, and the differences in scores are not reflected in differences in a measure of job performance, use of the selection procedure may unfairly deny opportunities to members of the group that obtains the lower scores" (Miner & Miner, 1979). This interpretation is easy to understand because the commonly used selection procedures often employ a scoring formula that combines the predictor scores into a single score on which the selection is based. In the development of ASVAB

composites, it is important that use of these composites in selection does not result in bias against any particular group.

There are a series of steps we can take in the present project in order to achieve this objective. First, we need to identify the factors that are related to group differences (with individual differences as the limiting case) in predictive validities and factors that may affect predictor and criterion reliabilities. We are concerned with the reliabilities of both predictors and criteria even though criterion reliabilities are traditionally considered as being less critical in validation efforts. This lack of emphasis on criterion reliabilities largely stems from the belief that measurement errors of the criteria do not affect the accurate estimation of the predictor-criterion relationship when Ordinary Least Squares (OLS) procedures are applied. This is not entirely true because such errors can lead to underestimates of predictive validity. More important, when the validation involves separate groups for whom the criterion has unequal reliabilities (as possibly the case in the present context), apparent group differences in validity estimates may be mistaken as true group differences in the strength of the predictor-criterion relationships. Moreover, if a common relationship exists among groups, failure to take into account the unequal reliabilities of the criterion in the OLS estimation procedure can produce inconsistent estimates of the relationship. Thus, in order to identify real group differences in the validation results, we have to be concerned with criterion reliabilities as well as predictor reliabilities. (It is well recognized that measurement errors of predictors can cause underestimation of the relationship.)

The concept of moderator variables is central to the technique employed to identify factors that may influence the predictor-criterion relationships. A variable is said to be a moderator if it does not have a direct relationship with the criterion but can influence the form and strength of the predictor-criterion relationship (Sharma, Durand, & Gur-Arie, 1981). A classical example is that predictability of freshman college grades tends to be higher for women than for men, so that sex is considered as a moderator (Saunders, 1956). This same concept has been generalized to include moderators for reliabilities (Ghiselli, 1963; Linn & Werts, 1971) as well. This usage is a direct extension of the preceding definition when we consider reliability to be a relationship between repeated measures of the same attribute (variable)--predicting the observed score from the true score. Furthermore, joint moderators which interact to influence the predictor-criterion relationships may also exist (Zedeck, Cranny, Vale, & Smith, 1971).

The relevance of moderating effects in the context of fairness in selection can be seen from an early definition of test bias: a test (predictor) is regarded as biased when the regression lines computed separately for the groups are different, because in that case, same scores on the test do not give the same predictions for members in different groups (e.g., Anastasi, 1968; Cleary, 1968; Guion, 1966). This definition implies that a test is biased (unfair) to individuals who differ in the moderator variable that accounts for the regression differences. Linn and Werts (1971) further point out two major ways that group differences can occur: (a) there may be differences in predictor-criterion correlations or in predictor reliabilities; (b) the regressions may differ in slope, intercept, or standard error of estimates.

It is important to emphasize that we do not subscribe to this early definition of test bias. A test may show different relationships to the criterion for members of different groups but still can be effectively employed in a fair selection procedure. We quite agree with Cronbach's (1976) distinction between predictive validity and appropriate selection policies, and consider a predictor battery to be inadequate only if it results in different selection (or classification) efficiency after the different relationships with the criteria among groups have been incorporated into the selection system.

Because predictive validity is likely necessary, though not sufficient, for effective selection, we deem it fruitful to investigate it by means of moderators.

There are several ways for identifying moderators that can be used to improve predictions. The three methods most frequently employed in empirical research are (Zedeck, 1971):

(a) Subgroup Analysis. Moderators are identified through comparisons of predictions for different groups (Frederiksen and Melville, 1954);

(b) Prediction of Predictability. Correlates of the absolute difference scores (the D scores) between standardized criterion and standardized predictor scores are identified and used as 'predictors of predictability' (Ghiselli, 1956; 1960); and

(c) Moderated Multiple Regressions. Multiplicative (cross-product) terms are introduced into the multiple regressions as new predictors (Saunders, 1955; 1956).

While all of these methods have been employed to examine moderating effects in validation research, the results using these techniques frequently differ. Further, moderators are not consistently substantiated in replications. This prompted Ghiselli (1972) to remark "it is possible that moderators are as fragile and elusive as that other will-o-the-wisp, the suppressor variable". We recognize the limitation of the moderator approach, but regard it as a useful concept in the present validation effort to support the development of a fair selection/classification system for the Army.

In our opinion, the difficulties in moderator research lie in the exploratory nature of the efforts. For this project, we shall take a confirmatory approach. We do not intend to engage in a fishing expedition to discover moderators. Instead, our research of moderating effects will follow explicit rationales. Specifically, sex and race/ethnicity are two variables we will examine closely for moderating effects because the law explicitly prohibits discriminations with regard to these characteristics. In addition, we will conduct a thorough review of the literature in personnel selection and consult with personnel experts in the Army and our research advisory panel to determine other candidate moderators and establish clear justifications for a need to examine them. (For example, in recent years, there seems to be growing concern about equal opportunities for children from bilingual and rural/urban backgrounds; there are also indications that high school graduation may also exhibit moderating effects, [Sims & Hiatt, 1981].) For these special-interest groups, we will carry out subgroup analyses in order to assure that there are no differential validities among the groups.

At the present, we anticipate that the potential moderators that we will investigate are to be qualitative (discrete) variables. Since natural groups can be formed in these cases, we can perform subgroup analysis in order to evaluate the effects of each potential moderator independently, and jointly if joint moderating effects are also suspected. Obviously, we may not always have sufficiently large samples to support the analyses. Where a particular group is specifically excluded by law from certain jobs (for example, women can not be assigned to MOS such as combat engineer and tank crewman), the issue of unfair selection for this group cannot be addressed. However, where assignment is permitted but there simply have not been a sufficient number of qualified applicants from some special interest groups, we will investigate alternative ways to conduct the evaluation of moderating effects for such cases. Because the selection of MOS for data collection in Task 5 has specifically taken sample sizes into consideration, we currently estimate that with few exceptions, we will have adequate samples to perform the analysis of simple moderator effects, at least for these MOS, but not necessarily for joint

effects. If preliminary analyses uncover evidence for important joint effects, we plan to obtain pooled estimates of such effects by combining some MOS.

Because assessment of validity is primarily based on the covariances between the predictors and the criteria (the criteria are assumed to have been put on comparable scales so that the differentiation is meaningful), we propose to perform an initial analysis employing the linear structural relation model (LISREL V, Joreskog & Sorbom, 1981) to ascertain the similarity of the predictor-criterion relationships among groups. (Linearity is assumed in the present discussion; if nonlinearity is suspected, we will consider polynomial models or transformations aimed at linearizing the relationships.) We will compare not only the variances-covariances but also the mean vectors, because mean differences can also lead to different regressions (i.e., intercepts). A set of hierarchical hypotheses can be tested to detect the exact form of differences among the groups. If differences do exist for one or more criteria, we will perform validation analysis for each group separately in order to select a best set of composites for each group or we will make adjustments of the composite scores for one of the groups. If group differences are not detected during the initial analysis, the particular characteristic being examined will not be treated as a moderator and will not be entered into further validations.

In practice, a moderator can function in several ways. Group differences may be found with respect to either the strength or the form of the relationships or both. In addition, a variable may bear a direct relationship with the criterion (and thus be useful as a predictor by itself), while at the same time influencing the relationships between other predictors and the criterion. Furthermore, if measurement errors are present, a difference in strength may be explained by differences in reliabilities or in true relationships or both. Thus, it is possible to distinguish among various types of "moderators" (Sharma et al., 1981).

The results of these moderator analyses will be applied to the computation of ASVAB composite (area) scores that are now the bases for selection/assignment of recruits to military services. Development of group-specific composites may be required in order to improve the efficiency of the ASVAB as a selection instrument. If such composites are obtained, we will have to compare their utilities with those of the earlier composites, so that appropriate recommendations on whether to use separate composites for special interest groups can be made. On the basis of previous validation (e.g., Maier, 1981), we do not anticipate to find consistent group differences to warrant use of group-specific composites.

Step 3: Determine the appropriate number of area composites.

The number of area composites to be defined will be determined by the extent of validity generalization across MOS. Accordingly, we will cluster the MOS on the basis of the similarities between their relationships with the predictors. The clustering will be accomplished in three computational steps.

(1) Compute predictor profiles.

We plan to use linear regression to produce for each MOS a "profile" of predictors of performance in the MOS. This profile would be a vector of standardized regresion coefficients, including both the ASVAB and specified additional measures, such as level of education.

(2) Eliminate unreliable cases.

Before proceeding, we plan to examine the results, MOS by MOS, to eliminate any outliers or any profiles that appear to be largely a function of error components. This will involve not only direct examination of the computer plots of the residuals but also comparison of the results with known and hypothesized sources of error variance.

(3) Hierarchically cluster the profiles.

We plan to use a procedure such as the SAS PROC CLUSTER in order to identify which MOS have similar profiles of performance predictors. We will examine the clusters produced for varying specifications of the number of allowable clusters, to determine the benefits to be gained by adding each additional cluster. Initially, we will obtain nine clusters to check against the nine MOS groups for which current composites are defined. Because the current MOS groups have been in use for some time and they are formed on the basis of expert judgments, any differences we found between our results and these groups will be carefully analyzed to explain the nature of differences. Additionally, we will also compare our clustering results with other clusters obtained on the basis of job/task requirement analysis such as that being

performed by RCA or that being investigated in Task 5. If substantial differences exist, we may have to reconcile the clusters by trying other methods. The number of clusters for the best solution will be taken as the number of composites to be defined in order to adequately differentiate between the MOS.

Step 4: Define the area composites for the MOS groups and set the cutoffs.

The mean profile of predictors for each cluster is the best overall predictor of performance in that cluster and as such satisfies the maximization requirement for the Area Composite. The beta weights will be transformed to integer weights, preferrably unit weights in keeping with the traditional practice. The resulting composites will then be evaluated in terms of their validities for each individual MOS within the group. If these validities are approximately equal to those obtained by individual profiles, we will suggest adoption of the composites and cutoffs for selection into the individual MOS (see later for a discussion of setting the cutoffs).

Computation of the mean profile for a cluster will be weighted by the size and, if available, importance of the MOS, and MOS that are close to the boundary between two clusters will be considered for logical removal to a different cluster. Furthermore, if odd clusters of MOS emerge, the hypothesis will be entertained that it is the deficiency of outcome measures that led to a distorted placing of the MOS. It is important for the Area Composites to be credible, or the likelihood of their proper use will be low; so all unusual placements will be carefully considered.

Careful comparison of the results with prior research and with current and past procedures is critical. It is essential that change not be instituted just for the sake of change, because there are substantial costs, particularly in terms of training for use of new procedures, that must be factored into the solution. On the other hand, we will have the opportunity of measuring the amount of information lost by forcing clusters, rather than providing a separate predictor vector for each MOS. This will provide one small part of the justification for the more comprehensively designed and validated enhancements to be developed during the remainder of the project.

As discussed earlier, it is important that the defined composites are fair to all groups when used in selection and classification. The area composites will be refined, if necessary, in order to insure their validities for various groups. Throughout the subtask, care will be taken to preserve the capability of analyzing results separately by race, sex, and other key variables. In particular, once clusters are defined, the predictors will be estimated independently for individual groups, in order to identify MOS for which the selected Area Composite is relevant for one group but not another. Should we find substantial effects, we will present these results to advisors who can evaluate the implications of various approaches to adjusting for the differences.

Although the DCSPER has the final authority to decide on the cutoffs for selection into an MOS, we will examine the effects of alternative cutoffs on

the selection outcomes in order to recommend appropriate cutoffs to the Army. These cutoffs will be calculated in terms of the trade-offs between failing to find qualified applicants versus failure to eliminate future wash-outs. If possible, we will propose to calculate multiple cutoffs that can be used upon notification of a change in urgency for filling an MOS.

Step 5: Conduct cross-validations.

Finally, cross-validations will be carried out to determine the extent to which the results represent real variation in ability requirements among MOS, as opposed to chance variation or artifactual variation in criterion measures.

In cross-sectional research, cross-validation is accomplished by dividing the available sample of cases into one or more pseudo-replicate samples. The simplest design divides the available sample into two halves randomly, develops the equation on one half, and cross-validates it on the other. This is in fact the classical cross-validation design. When the available sample is sufficiently large to permit it to be done with adequate precision, the available sample can be divided into two pseudo-replicates and the calculations are carried out in both. In general, the division into the two pseudo-replicates is not done randomly, and thus care must be taken to see that the splitting results in half-samples which are comparable.

Snee (1977) has developed an algorithm for splitting the data and evaluating the comparability of the two half-samples. This algorithm, as implemented in a computer program called DUPLEX, divides the data into two subsets that cover roughly the same region of the predictor space. The predictor scores are standardized and orthonormalized, and the Euclidean distance between all possible pairs of points is calculated. The two points that are farthest apart are assigned to the estimation set. The two points that are farthest apart among the remaining values are assigned to the validation set. Then, the point that is farthest from the two points in the estimation set is added to the estimation set; the point that is farthest from the two points in the validation set is added to the validation set; the point now farthest from the estimation set is added to it, etc. The pattern of alternative assignments continues until all the points have been assigned to one of the two sets. The comparability of the two sets is evaluated by computing the ratio of the determinants of the inverse of their information matrices. Since each determinant is proportional to the generalized variance of the corresponding predictor space, this ratio will be approximately 1.0 when the two half-samples contain roughly the same amount of information.

Although the point has not been much discussed in the literature, it may be advantageous to define multiple sets of pseudo-replicates, and repeat the estimation and cross-validation process within each. Clearly, when the total sample size is large, there is a very large number of possible pseudo-replicates that could be formed, and a single pair may not represent an adequate sample. The objective is to obtain the greatest possible precision with the fewest possible sets of pseudo-replicates.

McCarthy (1976) has argued strongly that a "balanced" sampling plan is the most effective. In particular, he has argued for a strategy developed for the analysis of complex sample surveys known as Balanced Half Sample Replications (BHSR). His analyses of simple cases have shown that this method is substantially more efficient than the traditional split into random halves. His simulation results for more complex cases suggest further that the advantages of BHSR are maintained for the cross-validation of multiple regression equations.

Balanced Half Sample Replication appears to extract nearly all the information available from all possible pseudo-replicates. Since the cost of BHSR analysis is substantially less than the complete pseudo-replicates analysis, this represents a considerable advantage. In conjunction with the more effective of the "shrinkage" estimation formulas (e.g., Wherry, 1931; Olkin & Pratt, 1958), we plan to use the work of Snee to select and evaluate half-samples, and if the sampling design for troops is balanced, use BHSR to define multiple sets of half-samples for cross-validation.

Further, when data are collected over several months, it is possible to split the data at a particular point in time, and to use the data collected after that point to validate the model constructed from the first half of the data. This type of design is useful in evaluating the stability of relationships. As it will take several months to accumulate the data from the major cohort test, it is possible to apply a form of rolling

longitudinal cross-validation design to these data, and to cross-validate both within and across time. This will make it possible to separate the effects of instability from the effects of sampling error. Naturally, use of such a longitudinal design is predicated on the characteristics of the accession sample not varying greatly over time.

Where feasible, this approach will be used to insure that the results are generalizeable to accessions who enter at different times of the year. Longitudinal analyses of this kind are far more powerful and efficient than using longitudinal data to do repeated cross-sectional analyses.

Prepare Interim Report(s). We expect to perform this subtask jointly with ARI staff. We will provide weekly reports to document and inform ARI of our rate of progress, so that there will be no unexpected failures to deliver products. We expect to present two major interim reports: one in July, 1983, for the Review Committee, in which we will carefully examine the implications of the proposed approaches; and one in October 1983, in which we will indicate the range of modifications that will be needed in the area composites.

The results of the analyses will be carefully reviewed. An adequate amount of time will be allocated to ensure that any re-analyses as may be required by ARI can be comfortably carried out, prior to the March 31, 1984, ending date for this subtask.

Prepare Final Report. We will submit a final report at the end of March, 1984, which contains all information needed to judge whether the use of new area composites for enlisted personnel selection and classification is warranted.

Second Analysis of ASVAB Scoring Procedures

The first analysis of ASVAB scoring procedures, the identification and validation of possible new area composite scores using FY81/82 cohort data, will be completed early in 1984. At the same time, new data on the FY83/84 cohort will become available from other Project A tasks for the 19 MOS in our sample and work can begin on a second analysis of the ASVAB scoring procedures. Data on new predictor and performance measures will be collected on this cohort between June and September of 1985. This second analysis will begin to bridge the gap between the current optimization procedure and computerized optimization. The major ingredient in the new capability will be the existence of utility scales that allow cross-MOS comparisons of the value of good versus poor performance. It will no longer be necessary to perform assignments based on the assumption that all performance increments are of equal value.

This second analysis will be undertaken in two parallel designs, one of which replicates in part the previous analysis of area composites and the other of which uses a prototype of the CPAS optimzation algorithm. In this

way, it will be possible to examine thoroughly the implications of changing to the CPAS, in terms of both benefits and costs. At the same time, the partial replication of the earlier area composite effort will provide the basis for estimating the stability of the measured MOS ability requirements over time. The previous method for developing area composites will be modified to include the new utility scales, in both the hierarchical clustering algorithm and the computation of mean predictor profiles for each cluster. The way in which the utility scales will be included in the former is as a weight on the difference between profiles: two distant MOS should be less likely to be forced to share a common area composite if the importance of good performance in each is critical than if good performance is not critical.

In addition to these differences, the second ASVAB analysis will benefit from the availability of new criterion measures being developed by Project A. While the operationalization of the ASVAB scoring procedures will continue, at this stage, to be in terms of area composites, these new sources of information may lead to more valid definition of composites.

For this second analysis of the ASVAB composites, we will be validating ASVAB forms 11/12/13 by assessing the discriminant validity of alternative area composites. The analysis will be performed between December, 1986 and March, 1987. A final report will be submitted, with recommendations on possible revisions of the ASVAB composites and new cutoffs if required.

Section 2: Support for the Development of New/Improved Measures

This section describes the procedures for Subtasks 1.2 through 1.5 in the draft Integrated Master Plan. The research team carrying out each task possesses ample analytical expertise to achieve its objectives. Nevertheless, the design of Project A has concentrated analytical and data processing capabilities in Task 1; and in order to maximize the benefits of carrying out this effort as a single project, Task 1 staff will assist the analysts in other tasks to promote efficiency and a uniform level of statistical and psychometric sophistication.

This assistance is implicit in the Research Plans for the other tasks, and the time lines for the assistance correspond to the requirements of those tasks. In this section, we briefly describe the current plans for analytical support for each of the other tasks. Support for data entry, editing, and access for the LRDB is described in the Draft Longitudinal Research Data Base Plan. Many of the analyses of the field test and 83/84 concurrent and longitudinal first tour sample data will be accomplished under these support subtasks.

Support for the Development of Pre-induction Predictors (Task 2). As a part of the development of ASVAB area composite scores, Task 1 will identify MOS for which the current ASVAB has greatest and least validity, after appropriate data adjustments, in order to suggest areas in greatest need of

improvement. This information will result from the examination of the existing ASVAB composites described in the preceding section (Subtask 1.7). Preliminary results will be passed to Task 2 in August, 1983 for use in their planning for the development of new measures.

In planning for analyses to support the development of measures, Task 1 will assist in sampling design for the field tests, in the development of approaches for Computer-Assisted Testing (CAT), and in testing of assumptions about the comparison of construct validity of different predictors.

Task 1 will focus particular attention on the problems and potential of CAT and will attempt to provide staff of Task 2 with knowledge gained as a function of validation of the CAT forms of the ASVAB. Integration of CAT with tailored testing and with latent trait theory will require significant research efforts, and Task 1 will monitor the results of research on CAT being carried out elsewhere.

Field tests of new pre-induction predictors will be carried out in 1983/84 with the Preliminary Predictor Battery and in 1985 with the Trial Predictor Battery. Task 1 staff will enter the data into the LRDB and then merge these data with various criteria in order to identify those predictors that contribute the greatest increment to differential validity for MOS assignment. In order to meet Task 2 needs for creation of the Experimental Predictor Battery, preliminary results of the analysis of the FY83/84

first-tour concurrent validation date must be made available to Task 2 by December, 1985.

Task 1 staff will provide Task 2 with edited data for their analyses of the psychometric properties of the predictor measures and will assist in these analyses. Task 1 staff will similarly assist the other tasks in the analysis of the psychometric properties of the various criterion measures. Task 1 staff will then perform preliminary validation analyses to provide Task 2 with information on the contribution to overall predictive validity made by each predictor measure and information on any problems associated with the differential validity of particular measures across gender and race/ethnic groups. These analyses are mentioned in Subtask 9 of the Task 2 research plan. Our approach to specific issues in these analyses is described in the following section (Section 3) on the Validation of New and Improved Predictors of Army performance.

The support for Task 2 will be facilitated by the inclusion of staff at the site carrying out Task 2 (i.e., Minneapolis) as a part of the Task 1 staff.

Support for the Development of Training Outcome Measures (Task 3). The initial analyses aimed at development of ASVAB composites will be based on the extent to which ASVAB subtests predict existing training outcome measures in different MOS. In the course of these analyses, Task 1 staff will rely on information gathered by Task 3 in their evaluation of the existing measures (Subtask 3.2). In turn, the use of these data as

criterion measures will also identify strengths and weaknesses in current training outcome measures that will be useful to the staff of Task 3 in their evaluaiton of these measures.

In planning for analyses for Task 3, the staff of Task 1 will assist in sampling design and in the identification of particular confounding factors that affect training outcomes, including variations between training sites. In carrying out the concurrent validation on the FY83/84 cohort, Task 1 will compare Task 3 measures with those being considered in Tasks 4 and 5 in order to develop hypotheses about the structure of the criterion space. Similarly, the structural models tested in the concurrent validation will provide essential information on the use of training measures as predictors of subsequent performance. This information will be fed back to Task 3 staff to aid in their conduct of Subtask 3.8.

The proximity of the sites at which Tasks 1 and 3 are headquartered, in Palo Alto and Carmel, California, will facilitate communication between these Tasks.

Support for the Development of Army-wide Criteria (Task 4). In the course of analysis of FY81/82 Enlisted Master File data for the purpose of ASVAB validation, Task 1 will perform numerous descriptive analyses of existing computerized Army-wide criterion measures. This descriptive information, along with information on data quality gleaned from the editing of these data is being passed on to Task 4 staff to aid in Subtask 4.1. In addition,

Task 1 will identify representative samples of FY81 accessions for use by Task 4 in searching noncomputerized (microfiche) records and will enter the resulting information merging it with other available measures for analysis by Task 4 staff.

In planning for analyses for Task 4, the staff of Task 1 will assist in scaling and aggregation problems and will focus particular attention on the development of utility measurement procedures. The translation of performance outcomes into a single utility scale which is comparable across MOS is vital to the derivation of the prediction equations that must be passed to Project B if CPAS is to become a reality. At the same time, the development of appropriate utility scales is one of the most methodologically complex tasks in this project. The data necessary for the development of utility scales will be collected under the auspices of Task 4 by Task 5 field personnel. Task 1 staff will input to the planning of this data collection effort and will support analyses of the resultant data to derive utility scales that are comparable across MOS. Task 4 will perform the difficult and critical task of preparing appropriate stimuli for scaling that accurately reflect the domain of performances to be scaled and preparing instructions for eliciting valid judgments of relationships among the stimuli. A description of the procedures for developing accurate stimuli and a valid judgment paradigm is given under Subtask 9 of Task 4.

Field tests of new first-tour Army-wide criteria will be carried out in 1984 and 1985, and Task 1 will assist in the incorporation of the data resulting

from these field tests into the LRDB, the analysis of the relations of these measures to MOS-specific measures, and the analysis of the different ways these two sets of criteria relate to measures taken at induction or during training.

Second-tour Army-wide criteria will be field-tested in 1986, and these data will be analyzed to identify predictors that should be taken into account at the time of making reenlistment decisions.

The coordination of efforts of Task 1 and Task 4 will be facilitated by the location of some staff of both tasks in the same office in Alexandria, Virginia.

Support for the Development of MOS-specific Criteria (Task 5). An initial effort confronting Task 5 has been the selection of MOS for special criterion development that will be representative of the entire range of MOS in the Army. The strategy has been to identify clusters of "similar" MOS, so that MOS can be sampled in such a way that every Army MOS is in the same cluster as at least one of the sampled MOS. As a part of the ASVAB area composite score development subtask, Task 1 will identify clusters of MOS with similar profiles of pre-induction performance predictors. Task 1 will also perform cluster analyses of MOS job content and requirement ratings. The results of these analyses will either confirm the adequacy of the MOS sample initially selected or suggest refinements of it.

Planning for Task 5 analyses will include sampling design; selection of jobs, tasks, and duty positions within MOS to maximize the generalizability of the results; and the design of administration procedures to support estimation of reliability and validity of MOS-specific criteria. Task 1 will assist in this effort in response to requests from Task 5.

As described under Activity 5.5.5, Task 1 staff will assist Task 5 in the analysis of field test data. A major component of this assistance will be in the implementation of LISREL V models for assessing construct validity and criterion equivalence. This approach involves the confirmation of constructs hypothesized to underlie the observed measures and the assessment of possible measurement bias from the use of different types of instruments. Such models will be used on a larger scale in the concurrent validation on the FY83/84 cohort to assess the extent of criterion equivalence across the Task 4 and 5 (and to a lesser extent Task 3) measures.

The coordination of efforts between Tasks 1 and 5 will be facilitated by the location of some staff of both tasks in the same office, in Washington, D.C.

Subtasks 1.2 through 1.5 described above provide Project A with the flexibility needed to respond to unexpected developments and problems in various areas. While focusing our effort on the validation of existing measures and the longitudinal predictive validation of new measures, we can allocate resources to assist one or another of the instrument development

tasks, as the needs arise. For example, programming support may be provided to Task 2 for the development of computer-mediated psychomotor tests.

At the same time, by maintaining careful documentation of files and project-wide availability of particular analytical procedures, Task 1 can eliminate many repetitive searches for similar data files and similar analytical packages. In this way, Task 1 will contribute to the cohesion of the project as a whole.

Section 3: Validation of New and Improved Predictors of Army Performance

This section presents the procedures for Subtask 1.8 in the Integrated Master Plan. During FY86/87, Project A will collect pre-induction measures and training data on a large sample of new recruits, and in the summer and fall of 1988, first tour Army-wide and MOS-specific performance criteria will be accumulated on this sample. In addition, second tour performance data will be collected for the FY83/84 cohort. The staff of Task 1 will analyze the relations among these measures in order: (a) to estimate the input parameters needed to drive the allocation optimization system produced in Project B; (b) to provide the empirical justification for the value of the project-developed selection/classification measures; and (c) to identify procedures for periodically updating the prediction system to maintain its integrity over time.

At this point, specific research plans for the longitudinal validation must clearly be viewed as tentative. It is difficult to predict the form which Army selection and classification procedures will take in 1986, and we shall have addressed a variety of methodological issues that may significantly alter the scope of our proposed analyses. Nevertheless, we can propose a series of steps that would guide our analyses if they were based on the current state of Army personnel procedures and of our knowledge.

The subtask can be divided into three efforts: (a) validation of the pre-induction measures used for initial MOS assignment; (b) validation of the training measures for use in re-assignment following training; and (c) preparation of inputs for the CPAS. The second tour FY83/84 cohort performance data including final first tour attritions, eligibility, and reenlistment rates will be analyzed along with the first tour FY86/87 data. However, we will not include the second tour performance measures for the FY86/87 cohort at this point, because they will not become available within the current time-frame for this subtask. A fourth analysis involving the use of first-tour performance measures in determining re-enlistment and attrition rates for the FY86/87 cohort will be made in 1991 under combined task efforts.

<u>Validation of Procedures for Initial Selection and MOS Classification</u>

This subtask will be carried out in ten steps, over a 54-month period, starting in April, 1985. The ten steps are as follows:

(1) Set objectives;

(2) Determine the data collection and cross-validation designs;

(3) Evaluate and revise objectives, if necessary;

(4) Acquire, check, and clean the data;

(5) Carry out preliminary analyses;

(6) Plan and execute the main validation analyses;

(7) Evaluate the cultural fairness of the proposed procedures;

(8) Prepare interim reports;

(9) Make final inputs to Project B allocation algorithm;

(10) Prepare the final report.

The following sections describe our present thinking with regard to each of these steps.

Step 1. Set objectives.

The objectives must be thoroughly and operationally defined in advance of the troop requests, so that an adequate data base will be assured. Therefore, the plans presented here must be updated, starting early in 1985, so that they can be reviewed by ARI with sufficient time for revision and refinement before presentation as part of the justification for the troop support request in May, 1985. (This troop support request will cover the administration of the Experimental Predictor Battery to the sample [possibly revised] MOS in the FY86/87 cohort.)

Step 2. Determine the sampling and cross-validation designs.

Unlike the earlier validation of the ASVAB using existing cohort data, this validation will require the administration of new instruments to a sample of recruits. As a result, the data base will have substantially fewer members,

although, we hope, the data on each recruit will be substantially richer and the numbers within selected MOS will be comparable to the numbers available for the earlier work. It is essential that adequate numbers of women and minorities be included, so that the effectiveness of the CPAS for each group can be evaluated.

A average of 2,200 trainees from each of the 19 focal MOS will be administered the experimental predictor battery as they enter AIT. The numbers administered the first-tour performance measures will be about 25-40 percent of the initial sample size because only 12 to 15 Army sites will be visited to collect first-tour performance data. In addition, a sizeable percentage of enlisted personnel will either have left the service or be otherwise unavailable for performance testing during the site visit period. As in the case of the FY83/84 cohort, enlisted personnel in the FY86/87 cohort who reenlist will be administered second-tour performance measures.

Prior to the analyses, the balanced half-sample pseudo-replicates will be defined. They will be used in the main validation analyses for the purpose of cross-validation.

Step 3: Evaluate and Revise Objectives, if Necessary.

Based on the troop availability and the budget for test administration, it may be necessary to limit the scope of the validation. On he other hand, if the results of the MOS clustering, FY81/82 file data simulations, and

similarity scaling are promising (see Validity Generalization section, pages 26-31), it may be possible to expand the analyses to simulate the application of the algorithm to a larger set of MOS. The impact of MOS subgroups heterogeneity on the applicability of regression equations will also be explored using the FY86/87 data (see page 29). The results of these as well as other analyses could lead to a revised set of objectives based upon what the preliminary analyses indicate can be accomplished with the data. In this regard, the collection and analysis of file data from other MOS (in addition to the 19) may be deemed warranted.

Step 4: Acquire, check, and clean the data.

Data will be acquired, edited, and entered into the LRDB in accordance with the procedures outlined in the Longitudinal Research Database Plan. The data will be checked for duplicate records and obvious keypunch errors, screened for outliers and implausible values using relational edits, imputed using a statistical algorithm, PROC IMPUTE, where necessary, and entered into the LRDB. Further details on our editing procedures are found in the LRDB Plan.

Step 5: Carry out preliminary analyses on predictors and criteria.

Basic descriptive statistics on the predictors and criteria will be computed. From this information, corrections for restriction of range and unreliability (where appropriate) will be computed using the methodology

discussed in the section on the development of ASVAB composites. After corrections, the appropriate descriptive statistics will be recomputed (e.g., correlations after restriction of range correction). Composite scores will be added to the file where appropriate.

Step 6: Plan and execute the main validation analyses.

The longitudinal validation will synthesize all that has been learned in the course of the project, both in terms of new measurement procedures and in terms of methodological issues. We envision an analysis strategy consisting of a series of eight steps, as follows:

Step 6A: Express the criterion measures in terms of expected utility to the Army.

The measurement of the differential utility to the Army of good versus poor performance on the critical tasks in each selected MOS is an essential component of the development of the CPAS. The only way in which assignments can be optimized is through comparison of the "payoff" to the Army for having the recruit assigned to this or that MOS. Payoffs resulting from good performance must be measured on the same scale for each MOS, and it is that scale, or numerical assignment, that we are referring to as "utility." Task 1 will translate the utility data, collected by Task 5 personnel, into the common scales needed for the selection and classification system and for

the validation effort, following the procedures developed in Subtask 4.9. MOS utilities for second tour performance levels will be applied to the FY 83/84 data.

Step 6B: Determine the conditions under which the validation analyses will be run.

The optimization algorithm, or an approximation to it, will be used to evaluate the expected results of computerized allocation. This algorithm can then be executed under a variety of conditions to test the sensitivity of the algorithm to different factors. These include: (a) omitting a subset of the predictor measures, (b) changing particular recruiting strategies (such as bonuses), and (c) changing relative utilities in different MOS. In the main validation analyses, the effect of these conditions on classification efficiency will be determined. Predictors will be evaluated in terms of increments in expected utility.

Step 6C: Find linear (and perhaps non-linear) combinations of predictors which maximally predict expected utility.

Both linear and non-linear scoring procedures for the predictor instruments will be sought which will maximally predict the utility criteria. An expected value will be obtained for each recruit in each MOS, and these will be aggregated as appropriate for use in the Project B algorithm.

Step 6D: Find Bayesian (or Empirical Bayes) combinations of predictors.

Bayesian simultaneous estimators often show greater stability than least square estimators (which maximize predictability in the sample). They also tend to have smaller mean square error, particularly when prior information about small subgroups can be incorporated into the estimates.

Step 6E: Compute the coefficient matrix for each condition (as defined in Step 6B).

We shall then compute the aggregate expected utility estimate for each condition. The results of various alternative optimizations will be compared with each other and with the best current procedures for selection and classification. Using either linear programming or the method agreed upon at that point in time, we shall compute the increments in utility that accrue as a result of each addition to the prediction system.

The battery will be evaluated in terms of classification efficiency by implementing the allocation algorithm to simulate training/job assignments for the FY86/87 recruits. The improvement in classification efficiency as a result of basing decisions on the CPAS serves as the final justification for its full-scale implementation. In order to ensure the long-term effectiveness of the CPAS, the validations will be repeated by varying the system parameters to represent a number of plausible conditions under which the CPAS will be applied.

Clearly, this evaluation of the practical value of the CPAS needs to also take into account the cost of its implementation. The intermediate validation process and the cost-benefit analysis in the course of the predictor development are aimed at insuring the practical utility of the CPAS. Nevertheless, at this final stage of validation, we shall make an integrated assessment of the system by simultaneously considering the cost and the gain in utility.

Step 6F: Determine the extent to which the inclusion of training measures improves classification efficiency.

This validation will closely parallel the preceding plan, adding training measures to pre-induction measures as predictors and using first-term performance as criterion. These analyses will address the costs and benefits of reclassification following training. The need for this validation will be determined in the course of the project, as the potential value of improved prediction for assisting in assignment decisions following training is clarified. This value will depend on a complex combination of scheduling constraints which limit the flexibility of reassignments following training.

Step 6G: Evaluate the adequacy of the training and Army-wide criterion measures as stand alone criteria.

To enable the Army to update the system feasibly, it is important that proxy measures for the detailed and expensive MOS-specific performance measures be

shown to be valid when used without the specific measures. Both training outcomes and Army-wide criteria will be used in the algorithm to determine the extent to which different personnel allocations would occur if validity parameters were based solely on criterion measures that did not include MOS-specific performance measures. In addition, the impact of using the less expensive MOS-specific measures (ratings and job knowledge tests) but not hands-on measures on the allocation outcomes will also be examined.

<u>Step 6H: Carry out the cross-validation and stability analyses.</u>

Task 1 will carry out the main validation analyses in accordance with the balanced half-sample cross-validation design chosen in Step 2. This involves repeated calculations in the half-samples specified in the design of the longitudinal validation. In addition, stability analyses, which are described in detail below, will be done using both the FY83/84 and FY86/87 databases.

It would be desirable if the predictor-criterion relationships could be shown to be stable over changes in time and conditions. Then the findings in the current validation could be continuously applied to provide the basis for predictions of job performance, while the supply requirement changes would be accommodated by modifying the constraints in the allocation algorithm. It is therefore important that we investigate the extent to which predictive relationships and utility structure change over time and conditions. Moreover, we need to determine the related factors causing the

changes so that we can anticipate the changes and plan for appropriate modifications to the allocation system.

The stability of the factor structure and interrelationships among measures will be assessed by comparing the analytic results between FY83/84 and FY86/87 cohorts. With regard to the stability of utility measures, Task 4 plans to repeat the data collection on utilities from a different group of Army staff and at different times (Summer of 1985 and 1988). These utility data will be analyzed to assess the stability of the estimation procedure.

We propose to conduct the stability analysis in two ways. The relationships will first be examined with conventional statistical methods in terms of means, and covariances. For the stability of factor structures, various stability coefficients have been suggested in a slightly different context -- that of factorial invariance across populations, and across measurement instruments/tests (Lawley & Maxwell, 1971; Meredith, 1964; and Mulaik, 1972). These indices can be extended to the present problem. However, we plan to address the issue of stability primarily by analyzing the data from the FY83/84 and 86/87 cohorts simultaneously with the LISREL model (Joreskog, 1971; McGaw & Joreskog, 1971; Sorbom, 1974; 1978; 1981; and Joreskog & Sorbom, 1980).

The simultaneous analysis of multi-sample data allows us to examine in detail how the relationships and measurement structures differ between samples. Not only we will be able to determine the degree to which the

variances-covariances and means differ across samples (e.g., population over time), but we will also be able to test the stability of regression coefficients directly with the model. Here the important question is whether analyses of two different samples leads to the selection of the same tests and weight them in substantially the same way for each cluster of MOS. If we find that the relationships differ, we will search for the explanation of these differences keeping in mind the concurrent versus predictive nature of the two cohort validations. The factors that are related to the changes of relationships will be identified so that updates of the prediction models and/or the classification instruments may be planned.

While the examination of the predictive relationships and measurement structures are informative because they constitute the foundations of the allocation system, the direct test of the stability of the performance of the CPAS rests on actual implementation of it under varying conditions. For this purpose, we will further carry out the evaluation of stability by implementing the allocation algorithm with different prediction models obtained with different components of the available data bases and compare the assignment outcomes in terms of classification efficiency. In conducting this evaluation, we will also simultaneously vary the system parameters to examine the effects on the allocation outcomes of changes in priority, political decisions, and social/ economical conditions (some of these changes will require modifications of specific constraints, some may simply require changes in utility structure).

Again, where the changes in relationships and/or system parameters appear to affect the performance of the allocation system, we will attempt to identify factors that account for such effects so that we can estimate the frequency with which the CPAS will require updating. In making this estimate, we shall also take into consideration the costs of updates and the benefits to be gained from the resulting better allocation outcomes.

These stability analyses will produce a summary of the findings with special emphasis on those relationships that are most likely to change with time and social/economical conditions, and the factors that are likely to cause such changes. We will also make recommendations on the update frequency for the CPAS and ways of updating the system.

Step 7: Evaluate the cultural fairness of the classification procedures.

The differential validity and reliability across racial and sex groups (and maybe other groups of special interest) will follow the procedures described earlier in conjunction with the validation of existing predictors (Section 1). However, the ultimate objective of the present endeavor is to insure cultural fairness with reference to classification efficiency. The investigation of moderating effects employing the concept of differential validity and reliability is a screening step. Having identified variables assignment outcome appear to have significant moderating effects on predicting differential job performance and having obtained estimates of the moderated relationships between predictors and criteria, our next step is to

apply these results to making assignments of individuals to jobs. The results of such assignments can then be evaluated for the effectiveness of employing moderators to assure fairness of a selection/classification procedure. That is, we want to identify for blacks as well as whites, and for women as well as men, those particular MOS where they can expect to make the greatest use of their abilities to contribute to the Army's mission and thus to progress most rapidly in a career ladder.

The same assignment procedure but using separate prediction equations (i.e., using different scoring formulas or different predictor sets) will be applied to each of the groups that are of special concern. Classification efficiency will be estimated from the assignment outcome and compared with the outcome obtained without consideration of different predictions across groups. In doing this, we will mostly conduct simulations using the available data base and assignment procedure. If for some reason the data base is not adequate in supporting such simulations, we will supplement it with computer generated data.

As the use of moderators is expected to improve predictions in general, we are interested in examining the accompanied improvement of overall classification efficiency for the applicant population as a whole. That is, we will compare the assignment outcome using moderated relationships with that using common relationships for all groups. Because the introduction of moderators into the allocation system incurs additional costs both in data collection and in algorithmic labor (computations and data management etc.),

we need to evaluate the practical value of each moderator for its ability to improve the fairness as well as the overall efficiency of the system. If incorporation of the moderated predictions into the system raises neither the degree of fairness among groups nor the total efficiency, then either we need to take a different approach to investigate the problem of fairness or we may have to accept the system as being appropriate (fair and efficient) within the limitation of currently available technology.

Step 8: Make final inputs to the CPAS.

The implementation of the CPAS being developed in Project B requires data from Project A concerning the prediction of performance in various jobs for each enlistee and the utility measures (effective coefficient) of the various performances. The exact form of the data to be input into the allocation system has to be specified by Project B. The staff of Task 1 will coordinate with the staff of Project B to determine what kinds of data are most directly useful to the system and in what format they should be provided. Then we will generate the required data according to the specifications.

The prediction of performance is to be based on the predictor-criterion relationships as estimated in the validation of the classification battery. The utility measures are to be based on the analysis of the value and preference judgments collected from the Army staff. Thus the required data

can be generated as by-products of tne validations in Task 1 and the analysis of utility data in Tasks 1 and 4.

At the present time, we believe that the most convenient form of input is to provide the classification measures for each enlistee to be assigned, a probabilistic prediction model for each MOS, and a utility function of the performance measures for each type of job (or in the case where a continuous utility function is not determined, a vector containing utility for each specified level of performance in that job). In this way, the computations of the expected utilities for each job-person match can be performed within the allocation system. The advantage of this form of inputs is that if any of the three elements that constitute the bases of utility computations changes (as may often be the case when social or political conditions change), the system can be operated without modification except for update of that particular data element(s).

An alternative would be to calculate the expected utilities for each of the possible job-person matches for direct inputs into the system. This may require less computational effort in the CPAS but will sacrifice the flexibility afforded in the first form.

In addition, depending on the actual design of the system, other data may be required for its operation. Based on our understanding of the proposed approach to be taken in Project B, the planning module of the system intends to treat individuals as members of classes and we are to develop the class

definitions for that use. If this is indeed the chosen strategy, we will define homogeneous classes of individuals on the basis of the classification measures so as to minimize the loss of information on the predictor-criterion relationships as a result of such grouping. The definition of these classes will depend in part on how sensitive the utility measure is in differentiating the various levels of performances that can be predicted from the classification measures. Clustering methods will be employed to divide the profiles of expected utilities of a representative sample of enlistees into a reasonable number of classes. Each class is then identified by the range of the predictor values for the group of enlistees in it. Then analysis of the predicted performances (in the utility scale) based on the class definitions and those based on actual classification measures will be conducted to assess the loss of information from using the class definitions.

A third category of data that may be requested by the CPAS is estimation of probabilities that a particular enlistee will survive the initial training and an estimation of the time and cost involved in the training before a recruit is assigned to the job unit. Similarly, estimates of the probability that an enlistee will fail to complete the first tour can be computed. Reenlistment rates and second tour performance levels and associated utilities can likewise be estimated. These findings, based on the FY83/84 cohort data, may be incorporated into the optimization objective to improve the utility of the allocation outcomes because they represent yet another kind of value and cost to the Army. The purpose of examining the potential use of these data in making assignment is to evaluate whether

utilization of such information will substantially increase the utility of the allocation outcomes.

The estimation of these probabilities and costs will be based primarily on the existing data base and may require additional effort to merge historical data into the longitudinal data base. At later time, data on attrition, training and reenlistment will be accumulated for t FY86/87 cohort and added to the earlier historical data to update the estimates of such probabilities and training costs. The new estimates can be used in the field implementation of the CPAS if Project B decides that use of such information improves the utility of its outcomes.

Step 9: Write interim reports.

Interim reports will be prepared as various stages of the analyses are completed. Reports are anticipated summarizing our findings with reagrd to the following issues:

(1) The costs and benefits of adding new predictors to the existing pre-induction battery.

(2) The costs and benefits of using information obtained earlier in soldier's career (selection/classification test scores, training success, and Army-wide measures to make later administrative decisions (eligibility for reenlistment, second-tour assignment).

(3) The stability and generalizability of our validation findings.

(4) Methods of continued updating of the estimation procedures for the CPAS, using, for example, "proxy" criteria.

(5) The effects of using moderator variables in the allocation algorithm.

Step 10: Write the final report.

The final report will conclude the work of the project, but it will pertain not only to the validation of pre-induction predictors but also to the validation of training measures as predictors and to the preparation of inputs to the CPAS.

SUMMARY OF EXPECTED OUTCOMES

The overall objective of Task 1 is to validate the new selection and classification instruments that are being developed by Project A for the Army. The longitudinal validation to be carried out with the FY86/87 cohort data is aimed at producing concrete evidence to justify the operational use of the new battery in the CPAS being designed by Project B. Along with this validation effort, Task 1 will develop prediction models of a recruit's future job performance in his/her Army career, and empirically determine the expected utilities to the Army if he/she is assigned to each of the particular MOS. The predictions and the utility measures are the foundations of the personnel allocation system. The longitudinal validation will also examine the incremental benefit and the feasibility of sequential decisions in the allocation system -- initial selection at the time of induction followed by reassignment decisions at various choice points such as post-training and reenlistment, as information on the recruit is accumulated from the early period of his/her Army career.

Task 1 will also carry out a sensitivity analysis of the performance of the allocation system by implementing the CPAS under varying conditions (e.g., altering system constraints to reflect differing priorities in the Army's mission, changing the utility structures to represent changing social, economical, and political situations, and adjusting system parameters to

accommodate new or revised policy decisions). This analysis will also be directed at identifying factors that can affect the effectiveness of the new allocation procedures. By examining the factors that are most critical in determining the efficiency of the system, Task 1 will make recommendations on the frequency of required updates, the components to be updated, and the updating procedures so that the effectiveness of the CPAS can be assured at all times.

It should also be added that, in order to maintain a continuity of the validation efforts initiated during this project, Task 1 also expects to provide a basis for on-going validations of the classification procedures. For this purpose, the extent to which the three types of performance measures (training success, Army-wide, and MOS-specific) are equivalent to one another will be carefully assessed. On the basis of this assessment, Task 1 will determine the feasibility of continued validations using earlier or more easily obtained performance measures as substitutes for later, and/or more expensive, measures of performance. If judged feasible, recommendations will be made regarding practical procedures for conducting satisfactory validations, employing more readily obtainable measures and adhering to resource constraints.

In the course of achieving this ultimate objective, Task 1 will perform interim validations aimed at providing information to aid in the development of new predictors that will have high predictive validity and can be efficiently employed in the CPAS. In addition, Task 1 will provide technological

and analytical support for the development of reliable and valid performance measures. These performance measures are required for the validation of the classification instruments, but may also prove useful for operational evaluations of enlisted personnel in the future.

A more immediate outcome of Task 1 will be the ASVAB composite scores to be recommended for use in the current selection and classification procedure. One of the first endeavors of Task 1 will be to validate the existing predictors (primarily ASVAB, high school education, and biographical data). This initial validation effort will, at the same time, produce recommendations on the best set(s) of composite scores by evaluating the differential validity of alternative scoring formulas and considering the cost-benefit trade-off of adopting new scoring procedures in the current selection practices. The impact of various cutoffs for selection of the recruits into an MOS or MOS family will also be determined.

In order to accomplish these practical objectives, Task 1 will have to investigate the conceptual and methodological issues surrounding personnel research (specifically, with regard to validation and development of measuring instruments). These investigations are expected to produce practical solutions to technical problems such as how to adjust for restriction of range, and how to measure differential validity and classification efficiency. Such solutions will be developed both by adapting existing techniques and by devising new methods. These research efforts will not only result in technical advances that are of use in the

present project as well as in other personnel research, but they will also generate scientific information that will increase our understanding of the personnel selection problems.

The operational and scientific outcomes that can be expected from Task 1 are summarized below.

Operational Outcomes

(1) A recommended set of ASVAB composites and cutoffs for use in the Army's current selection and classification procedures. Use of these composites is expected to increase the effectiveness of the initial classification decisions.

(2) Feedback information to other tasks that will support the development of valid and reliable new predictor and performance measures for the Army. In addition to serving as criteria in the validations, the performance measures may also be used in future operational evaluations of enlisted personnel. Such information will be extracted from the validation analysis as well as from the statistical and psychometric analyses of the field test data.

(3) Practical procedures for continued validations of the classification instruments. These procedures will be developed in conjunction with the investigation of criterion equivalence among performance measures.

(4) A cost-effective classification battery that can be administered easily and efficiently at the MEPs and be readily used by the CPAS for making training/MOS assignments. This will be achieved by the cost-benefit analyses of competing batteries while they are being developed.

(5) A sequential decision framework that will enable the Army to make reassignments at various choice points (e.g., post-training, promotions, and reenlistment). Through such a sequential decision process, the Army can increase its efficiency in personnel utilization.

(6) A common utility scale across the MOS and the performance levels within MOS, and based on this scale, a composite of a recruit's expected total utility to the Army (training performance, general conduct, as well as task-specific performance). This utility composite will be used as the effectiveness coefficient required to drive the allocation algorithm in the CPAS.

(7) Accurate prediction models for estimating the recruit's future performance utility from the classification measures. These predictions will serve as a fundamental basis for an efficient allocation system that aims to increase the effectiveness of performance in the Army.

(8) Recommendations on procedures for updating the CPAS so as to ensure its continued ability to achieve optimal classification outcomes and thus the best utilization of the recruits by the Army.

Scientific Outcomes

(1) Improved or new technical solutions to the methodological problems that plague research on personnel decisions. Examples of such problems are restriction of range, measures of classification efficiency and differential validity, validity generalizations, cross validations, nonlinear prediction models, moderating effects, and stability of estimations.

(2) New or improved applications of psychometric methods in the development of measuring instruments.

(3) Advances in the technology of using computerized adaptive testing in information gathering for the purpose of decision making.

REFERENCES

Anastasi, A. Psychological testing (3rd ed.). New York: Macmillan, 1968.

Brogden, H. E. On the interpretation of the correlation coefficient as a measure of predictive efficiency. Journal of Educational Psychology, 1946, 37, 65-76.

Cleary, T. A. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Cochran, W. G. Some effects of errors of measurement in linear regression. Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, 1970, 527-539.

Cronbach, L. J. Equity in selection--where psychometrics and political philosophy meet. Journal of Educational Measurement, 1976, 13, 31-41.

Frederiksen, N., & Melville, S. D. Differential predictability in the use of test scores. Educational and Psychological Measurement, 1954, 14, 647-656.

Gillman, L., & Goode, H. H. An estimate of the correlation coefficient of a bivariate normal population when X is truncated and Y is dichotomized. Harvard Educational Review, 1946, 16, 52-55.

Ghiselli, E. E. Differentiation of individuals in terms of their predictability. Journal of Applied Psychology, 1956, 40, 373-377.

Ghiselli, E. E. The prediction of predictability. Educational and Psychological Measurement, 1960, 20, 3-8.

Ghiselli, E. E. Moderating effects and differential reliability and validity. Journal of Applied Psychology, 1963, 47, 81-86.

Ghiselli, E. E. Comment on the use of moderator variables. Journal of Applied Psychology, 1972, 56, 270.

Greener, J. M., & Osburn, H. G. Accuracy of corrections for restriction in range due to explicit selection in heteroscedastic and nonlinear distributions. Educational and Psychological Measurement, 1980, 40, 337-347.

Guion, R. M. Employment tests and discriminatory hiring. Industrial Relations, 1966, 5, 20-37.

Heckman, J. J. Sample selection bias as specification effort. Econometrics, 1979, 7, 153-161.

Joreskog, K. G. Simultaneous factor analysis in several populations. Psychometrika, 1971, 36, 409-426.

Joreskog, K. G., & Sorbom, D. Simultaneous analysis of longitudinal data from several cohorts. Research Report 80-5. Uppsala, Sweden: University of Uppsala, 1980.

Joreskog, K. G., & Sorbom, D. LISREL V: Analysis of linear structural relationships by the method of maximum likelihood. Research Report 81-8. Uppsala, Sweden: University of Uppsala, 1981.

Lawley, D. A note on Karl Pearson's selection formulas. Royal Society of Edinburgh Proceedings, Section A, 1943, 62, 28-30.

Lawley, D. N., & Maxwell, A. E. Factor analysis as a statistical method (2nd ed.). London: Butterworths, 1971.

Lindley, D. N. Regression lines of the Linear functional Relationship, Journal of the Royal Statistical Society, Series B, 1947, 9, 218-244.

Linn, R. L. Fair test use in selection. Review of Educational Research, 1973, 43, 139-161.

Maier, M. H. Validation of selection and classification tests in the Army. Working paper personnel utilization Technical area 82-2. Alexandria, VA: U.S. Army Research Institute, November 1981.

Maier, M. H. Issues for defining ASVAB 11/12/13/14 aptitude composites. Alexandria, VA: Center for Naval Analysis, 1982.

McCarthy, P. J. The use of balanced half-sample replication in cross-validation studies. Journal of the American Statistical Association, 1976, 71, 596-604.

McGaw, B., & Joreskog, K. G. Factorial invariance of ability measures in groups differing in intelligence and socio-economic status. British Journal of Mathematical and Statistical Psychology, 1971, 24, 154-168.

Meredith, W. M. The estimation of criterion parameters from a biased sample (Doctoral dissertation, University of Washington). Dissertation Abstracts, 1958, 19, 981.

Meredith, W. M. Notes on factorial invariance. Psychometrika. 1964, 29. 177-185.

Miner, M. G., & Miner, J. B. Analysis of Uniform Guidelines on Employee Selection Procedures. Washington, D.C.: Bureau of National Affairs Educational System, 1979.

Mulaik, S. A. The foundations of factor analysis. New York: McGraw-Hill, 1972.

Novick, M. R., & Thayer, D. T. An investigation of the accuracy of the Pearson selection formulas. Research Memorandum (RM-69-22). Princeton, NJ: Educational Testing Service, 1969.

Olkin, I., & Pratt, J. W. Unbiased estimation of certain correlation coefficients. Annals of Mathematical Statistics, 1958, 29, 201-211.

Rydberg, S. Bias in prediction. Stockholm: Alqvist and Wiksell, 1963.

Saunders, D. R. The "moderator variable" as a useful tool in prediction. Proceedings of the 1954 invitational conference on testing problems. Princeton, N.J.: Educational Testing Service, 1955.

Saunders, D. R. Moderator variables in prediction. Educational and Psychological Measurement, 1956, 16, 209-222.

Sharma, S., Durand, R. M., & Gur-Arie, O. Identification and analysis of moderator variables. Journal of Marketing Research, 1981, 3, 291-300.

Sims, W. H. An application of factor analysis to the constrution of improved classification composites from the Armed Services Vocational Aptitude Battery (ASVAB) Forms 6 and 7. Arlington, VA: Center for Naval Analyses, 1978.

Sims, W. H., & Hiatt, C. M. Validation of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 6 and 7 with applications to ASVAB Forms 8, 9, and 10. CNS 1160, Research report. Alexandria, VA: Center for Naval Analyses, 1981.

Sims, W. H., & Mifflin, T. L. A factor analysis of the Armed Services Vocational Aptitude (ASVAB) Forms 6 and 7 (Memorandum). Arlington, VA: Center for Naval Analyses, 1978.

Snee, R. D. Validation of regression models: Methods and techniques. Technometrics, 1977, 19, 415-428.

Sorbom, D. A general method for studying differences in factor means and factor structures between groups. British Journal of Mathematical and Statistical Psychology, 1974, 27, 229-239.

Sorbom, D. An alternative to the methodology for analysis of covariance. Psychometrika, 1978, 43, 381-396.

Sorbom, D. Structural equation models with structured means. In K. G. Joreskog & H. Wold (Eds.), Systems under indirect observation: Causality, structure, and prediction. Amsterdam: North-Holland, 1981.

Srinivasan, J., & Weinstein, A. G. Effects of curtailment on an admissions model for a graduate management program. Journal of Applied Psychology, 1973, 58, 339-346.

Wherry, R. J. A new formula for predicting the shrinkage of the coefficient of multiple correlation. Annals of Mathematical Statistics, 1931, 2, 446-457.

Wise, L. L., & Wang, M. Project A: Development of Improved Army Selection and Classification Systems Longitudinal Research Database Plan [Draft]. Human Resources Research Organization, 1983.

Zedeck, S. Problems with the use of "moderator" variables. Psychological Bulletin, 1971, 76, 295-310.

Zedeck, S., Cranny, C. J., Vale, C. A., & Smith, P. C. Comparison of "joint moderators" in three prediction techniques. Journal of Applied Psychology, 1971, 55, 234-240.

TASK 2 RESEARCH PLAN

PRE-INDUCTION PREDICTION OF ARMY SUCCESS

GENERAL PURPOSE OF TASK 2

The general purpose of Task 2 is to identify an efficient and effective set of initial or pre-induction predictors of soldier performance. By efficient, we mean that time and money to be expended on operational administration of the predictors is kept as low as possible, and by effective, we mean that the predictors forecast as accurately as possible the degree of success to be expected of recruits in various aspects of soldier performance, e.g., overall adaptation to the Army, completion of training, performance in specific MOS, and reenlistment.

There are two different, but related aspects to this general purpose. First, we will evaluate the effectiveness of the present set of initial predictors used by the Army. The major initial predictor now in use is a set of cognitive tests, the Armed Services Vocational Aptitude Battery (ASVAB). Prior research shows that this battery and its similar predecessors are fairly effective in predicting how well soldiers will perform during training, but there is much less information available about its effectiveness in predicting other important areas of soldier performance, notably on-the-job performance. Task 2 in conjunction with Task 1 will perform research to evaluate the effectiveness or validity of the ASVAB for predicting these additional aspects of performance.

Second, the ASVAB contains only cognitive tests. Measures of other types of human abilities and characteristics have been shown to be useful for predicting effective performance in training and on the job in a number of occupational areas (Dunnette, 1976; Owens, 1976). Task 2 will identify and develop new predictors that are most likely to be effective and efficient additions to the present set of predictors. The validity or effectiveness of these new predictors will be investigated in the same way as the validity of the present set of predictors. The evaluation of the efficiency of newly developed predictors will require analysis of the improvement in prediction of soldier performance gained by use of the new predictors over that obtained by the sole use of the present set of initial predictors.

Thus, the two aspects of Task 2 are closely intertwined: new predictors will be developed to add to and complement the present set of predictors; the validity of all predictors for all aspects of soldier performance will be scrutinized, and the complete set of present and new predictors will be analyzed to identify the most efficient set of initial predictors. Intertwined as well are the school-based criteria and predictors developed in Task 3, the Army-wide measures and utilities developed in Task 4, the specific-job measures developed in Task 5, and the integrated analyses of Task 1. The "one project" nature of Project A is perhaps best illustrated in the evaluation of existing and new initial predictors.

BACKGROUND ISSUES AND RATIONALE

Performance Relatedness: Construct Validation

Accuracy of personnel selection and classification decisions rests ultimately on the degree of congruence shown between performance scores on selection measures and performance measures collected some time later within the settings or jobs where persons have been placed. As Wernimont and Campbell (1968) have suggested in their widely cited article, congruence is most likely when elements of the predictor measures actually sample aspects of the performance domain to be measured, as opposed to resting on untested assumptions about relatedness between predictor performance and later performance.

The argument leads to a research strategy designed to identify patterns of behavioral consistency linking meaningful samples of performance drawn from different contexts, e.g., academic performance, extracurricular activities, job, and work performance, leisure pursuits, behavioral responses to standardized simulations, etc. Linkages between different areas may be made by logical analysis of activity components into directly measurable tasks and knowledge (content-oriented strategy), by rational analysis (by job and/or subject matter experts) of activity components according to relatively consistent behavioral abstractions (construct-oriented strategy), or by demonstrating statistical patterns of similarity across activity components (criterion-oriented strategy).

The behavioral consistency point of view has the following conceptual and methodological implications for the process of evaluating, identifying, and developing new pre-induction predictors of Army performance effectiveness:

1. Choice of domains within which to develop predictors is focused on evaluations of performance and performance relatedness. In the present context, the areas of performance to be investigated are general adjustment to a career in the Army, success in Army school training, effectiveness in specific Army tasks and jobs, and choosing to continue an Army career through reenlistment.

2. Validation of predictor measures begins with a content-oriented strategy that may be elaborated with existing behavioral theory (construct-oriented strategy) and confirmed ultimately with empirically demonstrated statistical relationships (criterion-oriented strategy).

3. An increasing number of scholars (e.g., Dunnette & Borman, 1979; Guion, 1980; Messick, 1980; Cronbach, 1980; Peterson & Bownas, 1982) is placing primary emphasis on the construct-oriented phase of the above strategy. Arguing, in effect, that all validation is really construct validation, they point out that validation encompasses: (a) a theoretical or conceptual component, the more highly developed the better; (b) a developmental step involving instrumentation based on content validation in which judgments are made about the appropriateness of content for specific objectives; and (c) an empirical component that entails determining empirical relationships between a

measure and other measures. Therefore, any discussion of the validity of a predictor for a specific purpose must summarize the best available information for all these components.

In sum, then, the approach to evaluating existing pre-induction predictors and in developing and evaluating new ones rests on behavioral consistency and construct validation as central concepts. Dimensions of soldier performance will be identified which are internally consistent and relatively distinct from each other. Corresponding classes of predictor measures will be structured on the basis of either direct behavior sampling or strong conceptual or theory-based inferences, and linkages between the performance dimensions and predictors will be evaluated empirically.

Validity Generalization and Situational Moderation

Hunter, Schmidt, and Jackson (1981) have presented methods they call State-of-the-Art Meta Analysis for use in helping to decide whether or not validity results may be generalized across different situations and populations. They present and illustrate methods for correcting observed sample validities for such artifactual components as restriction in range, criterion and predictor unreliabilities, and variability due to sampling error. Several recent investigations (Dunnette, et al., 1981; Pearlman, Schmidt, & Hunter, 1980; Schmidt, Hunter, & Caplan, 1981; Callender & Osburn, 1981; Schmidt, Hunter, & Perlman, 1981) of large data sets have shown that validities are more likely to be generalized across settings than traditional thinking in selection research has assumed them to be.

Nonetheless, some amount of variance in validities frequently remains even after the so-called artifactual components leading to variability have been taken into account. For example, Linn, Harnisch, & Dunbar (1981) report that about 30 percent of the variance across 726 validities between the Law School Admissions Test and first-year grades in law school remains after estimating and removing variance due to artifactual components. Dunnette, et al. (1981) estimates that 35 percent of validity variance across 70 electric generating companies may be moderated by situation or company specific factors.

Considering the accumulating evidence cited above, it seems reasonable to expect that validities are, in fact, much more general across different settings than has previously been presumed. We may reasonably estimate that the amounts of variance in validites due to situational components probably range between 0 percent and a maximum of perhaps 40 percent. In the context of predicting performance by enlisted personnel in the Army, it is our expectation that both existing and new pre-induction predictors will show validities that are general across quite broad groupings of Army settings, Army performance measures, Army population subgroups, and Army tasks and jobs. Our methodological strategy will, however, test empirically the limits of such validity generalizations and capitalize upon patterns of situational moderation where they are found to exist. In effect, by evaluating validity results in this way for both current and new pre-induction predictors, we will be able to discover patterns of similarity and differences between validities according to such potential situational taxonomies as types or families of predictor constructs, dimensions of

Army-wide success, clusters or families of Army schools, Army job (MOS) families, and sex and race subgroups.

Our strategy will be to discover predictors with validities that generalize across sex and race subgroups and across all dimensions of Army performance. However, we shall also seek to discover predictors that possess different validities across a sample of Army jobs in order to make more likely the increased efficiencies of personnel utilization which can be realized by strategies of differential placement or classification.

Efficiencies of Classification

Ideally, the Army would place all persons in jobs best suited to them, an outcome that would somehow assure not only that each person could use his/her abilities in the best possible way but that the Army could also allocate its human resources optimally across all available job assignments. This state of affairs is never perfectly attainable, but the goal of personnel classification in the Army should be to seek to optimize the matching of human skills with job requirements within the constraints dictated by the particular mix of human resources and jobs that exist at any particular time.

In fact, the numbers of jobs available in the Army and their great diversity afford opportunities for efficient utilization of human resources not available for most employers. Many of these potential advantages are

described by Dunnette (1966, pages 184-185). The most important advantage was recognized and described many years ago by Brodgen (1951). He showed that a personnel classification strategy could yield a marked increase in overall job effectiveness for an organization by making possible the use of more advantageous selection ratios for each of the distinct job areas to be filled.

In spite of the obvious advantages to be gained from classification as opposed to pure selection, classification procedures are often difficult to implement because of the many constraints operating in the real world of personnel decision making. In the Army, such constraints might include stated preferences (for job areas, types of schooling, geographic location, etc.) by recruits, special skill or educational requirements, the Army's ability to estimate human resource requirements and to provide appropriate training or job assignments over the long term as opposed to needs for filling immediate short-term vacancies, and the nature of cyclical fluctuations in recruit availability throughout the year. Most of all, such a strategy requires input of empirical information about the way in which the various aspects of soldier performance is, in fact, predicted by different configurations of scores on pre-induction predictors. It is crucial that attention be given to the development of such predictors from the beginning of the research effort and that all estimates of predictor-criterion linkages be evaluated against a backdrop of emphasizing differential predictor validities for differing assignments in the Army. This is to assure that the efficiencies of accurate classification procedures may be realized by the Army in its recruitment and training of enlisted personnel.

Incremental Validity and Utility

As noted in the statement of general purpose of Task 2, a major part of the activity will be devoted to estimating the amount of improvement in prediction obtained by combining new pre-induction predictors with current predictors, or estimating the amount of incremental validity.

Incremental validity has traditionally been expressed according to statistically significant increases in variance accounted for (R^2) as a result of adding more predictor components or in terms of statistically significant differences between product moment coefficients or hit rates derived from competing selection strategies, e.g., old versus new predictors; clinical versus statistical prediction, etc.

Task 2 will make use of such incremental validity comparisons, but these comparisons are intended as way stations on the road to developing more meaningful comparisons of the relative increase in utility or benefit attributable to the addition of new predictors. These incremental utility analyses will be conducted primarily by Task 1 and Task 4 staff on Project A and will consider costs associated with use of new pre-induction predictors and benefits associated with different levels of job performance. Results of these analyses will, however, inform decisions made in Task 2 about the relative usefulness of the newly developed pre-induction measures.

Impact of Computers

Technological advances in computer hardware, particularly the development of microcomputers, and the application of computers to psychological measurement have several implications for Task 2. First, testing can potentially be decentralized with much less risk to the security of scoring keys and a large gain in the standardization of administration. Second, measurement of human abilities or other characteristics can be accomplished in less time and with improved psychometric quality. Third, types of human abilities that were difficult to measure because of unwieldy, unreliable, or extremely costly apparatus are now much more feasible to measure. This applies especially in the area of perceptual/cognitive abilities. Finally, the advances in microcomputer hardware are so rapid that what was not feasible yesterday due to high cost is very likely to be feasible today, and almost certainly will be tomorrow. In light of all this, our approach to Task 2 will be to make as full use as possible of the capabilities of present computer technology, and to keep foremost in our view the strong possibility of ultimate computer administration of all newly developed pre-induction measures.

Expansion of the Measured Predictor Space

The present pre-induction predictors used for Army enlisted personnel (EP) selection and classification include the following: scores on the subtests of the ASVAB, age, sex, high school diploma status; and a biographical questionnaire, the Military Aptitude Profile (MAP). ASVAB 8/9/10 contains 10 short subtests of cognitive abilities and achievement. The subtests,

in various combinations of from three to five each, are used to form aptitude composites for use in predicting training school success in various specific MOS groups. Maier (1981) reports mean validity coefficients ranging from .26 to .52.

Maier (1981) also shows that ASVAB validities generalize, for the most part, across racial and sex subgroups. Results for ASVAB agree, therefore, with findings reported by Schmidt and Hunter and their colleagues (Schmidt, Hunter & Pearlman, 1981; Schmidt & Hunter, 1978; Schmidt, Hunter, Pearlman & Shane, 1979) who have presented convincing evidence that cognitive tests show empirical validities that appear to generalize quite strongly across a number of different training and job performance areas in the Army. However, it is important to note that their investigations, similar to the ones summarized by Maier, have so far been restricted to cognitive measures and that the degree of validity generalization is greater across training or school settings than across specific MOS task performance settings.

This set of predictors, dominated by the ASVAB, is somewhat limited when considered against the total range of tests, questionnaires, and inventories that have been developed by psychologists over the last 50 years. Current pre-induction predictors include no psychomotor, perceptual, or physical measures. Nor do they include measures of personality or vocational interest. Use of instruments from these domains would certainly increase the measured portion of the total possible predictor space, and thereby enhance the possibilities of obtaining improvements in the accuracy of selection and classification. The net effect of these considerations is

to impress upon us the importance of developing and evaluating a richer array of pre-induction predictors than have been used thus far in the assessment of Army EP candidates.

Rationale

The rationale for the conduct of Task 2 is founded on the consideration of the issues discussed above. Predictors from all the major domains of measured human attributes will be considered and considerable attention will be given to evaluating potential pre-induction predictors from the viewpoint of behavioral consistency and construct validation. A strong case must be made for a new predictor, based on its documented empirical relationship to job performance dimensions similar to those identified as defining soldier performance, or on its solid rational or theoretical relationship to such dimensions. Each predictor will be empirically evaluated with regard to its validity for a number of MOS (validity generalization), its unique contribution to selecting and classifying candidates (incremental validity and validity moderation based on MOS differences), and its present or potential degree of implementation via computer.

SPECIFIC OBJECTIVES

Objectives of Task 2 research are as follows:

1. Identify measures of human abilities, attributes or characteristics which are most likely to be effective in predicting, prior to entry into the Army, successful soldier performance in general and in classifying persons into MOS where they will be most successful, with special emphasis on attributes not tapped by current pre-induction measures.

2. Design and develop new measures or modify existing measures of these "best bet" predictors.

3. Develop materials and procedures for efficiently administering experimental predictor measures in pilot tests and to the FY83/84 and FY86/87 cohorts.

4. Estimate and evaluate the reliability of new pre-induction measures and their vulnerability to motivational set differences, faking, variances in administrative settings, and practice effects.

5. Determine the interrelationships (or covariance) between the new pre-induction measures and current pre-induction measures.

6. Determine the degree to which the validity of new pre-induction measures generalizes across MOS, i.e., proves useful for predicting measures of successful soldier performance across quite different MOS <u>and</u>, conversely, the degree to which the measures are useful for classification or the differential prediction of success across MOS.

7. Determine the extent to which new pre-induction measures increase the accuracy of prediction of success and the accuracy of classification into MOS over and above the levels of accuracy reached by current pre-induction measures.

OVERALL SUMMARY OF THE PROCEDURE

There are 15 procedural steps or subtasks in Task 2. Figure 2.1 shows a time table for the subtasks, and Figure 2.2 shows the relationships between the subtasks and the general nature of inputs required from other Project A tasks. Below, we briefly summarize each of the 15 subtasks. Subtasks are then described in detail in the PROCEDURE section to follow.

1. Literature search and planning. Civilian and military research about the relative "success" of predictors for purposes of personnel selection and classification will be searched. Results of the search will be organized to facilitate selection of a preliminary battery and formal technical and cost reviews of potential predictors.

2. Selection of preliminary battery and preparation for administration to FY83/84 longitudinal sample. A set of "off-the-shelf" predictors that comprehensively and efficiently cover the predictor space will be identified by Task 2 staff and reviewed by ARI. After approval, the predictors will be obtained and administration procedures prepared.

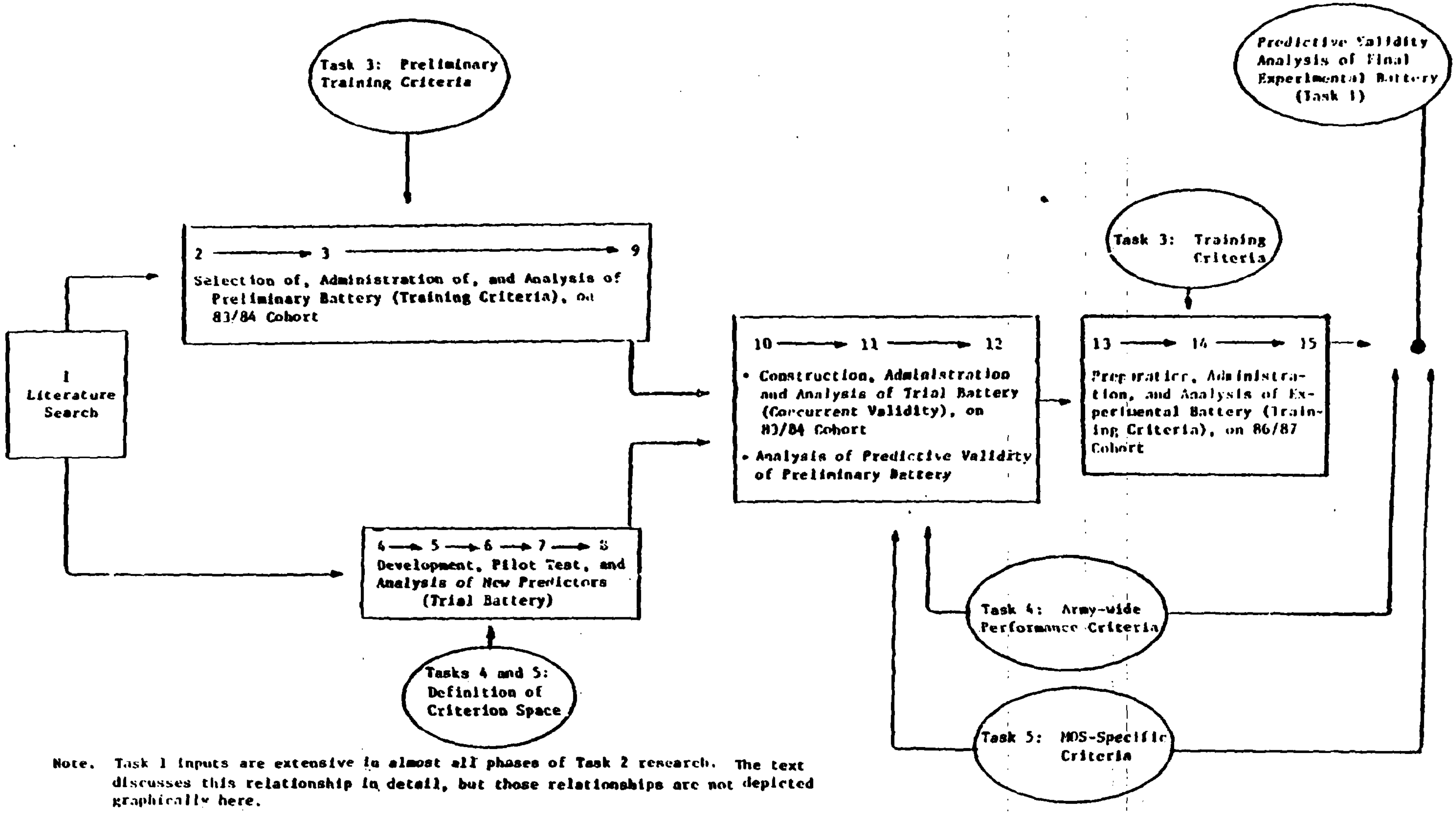
3. Administration of preliminary battery to FY83/84 longitudinal sample. The preliminary battery will be administered to a sample of 2,100-4,600 soldiers in training for each of four MOS: 05C, 19E/K, 63B, and 71L. On-site administrators will be trained by Task 2 staff and the administration process will be monitored by Task 2 staff.

FIGURE 2-1 TIMETABLE FOR TASK 2

Project Months and Calendar Periods: 1982, 1983, 1984, 1985, 1986, 1987 (months 6, 12, 18, 24, 30, 36, 42, 48, 54, 60)

Subtask	Start month	End month
2.1 Literature Search and Planning		7
2.2 Selection of Preliminary Battery	5	10
2.3 Administration of Preliminary Battery to 83/84 Cohort	12	21
2.4 Technical Review of Predictor Constructs and Measures	8	13
2.5 Cost/Administrative Practicality Review	13	14
2.6 Initial Development of Trial Battery	10	21
2.7 Pilot Tests of Trial Battery	22	24
2.8 Analysis of Trial Battery	25	27
2.9 Analyze Preliminary Battery	17	27
2.10 Prepare Trial Battery for 83/84 Cohort Predictor/Performance Data Collection	28	32
2.11 Monitor/Assist Administration of Trial Battery to 83/84 Cohort	33	36
2.12 Analyze 83/84 Cohort Data: Trial Battery/Performance Measures	37	39
2.13 Prepare Experimental Battery and Prepare for Administration of 86/87 Cohort	40	41
2.14 Monitor Administration of Experimental Battery, Further Analyses of 83/84 Cohort Data	42	53
2.15 Analyse 86/87 Cohort Data and Prepare Final Reports	54	60

FIGURE 2-2 RELATIONSHIPS OF TASK 2 SUBTASKS AND INPUTS FROM OTHER PROJECT A TASKS



Note. Task 1 inputs are extensive in almost all phases of Task 2 research. The text discusses this relationship in detail, but those relationships are not depicted graphically here.

4. Technical review of predictor constructs and measures. Experts will be used to make formal judgments about the usefulness of predictor constructs and measures for predicting soldier performance. Analyses of these judgments will identify a set of predictors judged to be "best bets" in terms of validity and efficiency, i.e., minimal overlap between predictors.

5. Cost administrative/practicality review. A panel of Army personnel knowledgeable about the field operation of recruitment, selection, and classification will be given information about the administrative procedures, costs, item types, etc., of the predictors surviving the technical review. They will make judgments about administrative feasibility, possible operations problems, etc.

6. Initial development of predictors for the Trial Battery (new predictors). Predictors chosen for development on the basis of Subtasks 4 and 5 will be designed and small scale tryouts will be held.

7. Pilot tests of Trial Battery (new predictors) in the field. The new predictors will be administered to a sample of soldiers from the FY81/82 cohort and a sample of applicants. Data will be collected to allow the investigation of practice effects, fakeability, motivational effects, and stability of measures.

8. Analysis of Trial Battery. The pilot test data will be analyzed to investigate administrative problems, applicant acceptance, psychometric properties, fakeability and practice effects, and covariances

of new predictors, current pre-induction predictors, and any criterion information available for the soldiers in the sample.

9. Analyze Preliminary Battery: FY83/84 cohort school and preliminary battery data. The covariance of the predictors in the Preliminary Battery with current pre-induction predictors and training success will be analyzed. Also, analyses of differences in MOS scores on constructs measured by the Preliminary Battery will be analyzed.

10. Prepare revised Trial Battery for FY83/84 cohort predictor/performance data collection. Information from Subtasks 8 and 9 will be integrated and plans formulated for revising the new predictors. After review and approval by ARI, revisions of the new predictors will give rise to the revised Trial Battery. Training will be provided to Project A staff responsible for administering the battery.

11. Monitor/assist administration of Trial Battery to FY83/84 cohort. Although the major burden of trial battery administration will be borne by other Project A staff, Task 2 staff will administer the battery for test-retest purposes and to a sample of new recruits, as well as providing "on-call" assistance for the major administration effort.

12. Analyze FY83/84 cohort data: trial battery/performance measures. Data will be available for concurrent validity analyses of the Trial Battery and predictive validity analyses of the Preliminary Battery,

although on fewer MOS for the latter than for the former. Fairness analyses, generalizability analyses, and other analyses will also be conducted.

13. Prepare Experimental Battery and prepare for administration to FY86/87 cohort. Based on analyses from Subtask 12, the Experimental Battery, i.e., the final, revised version of the Trial Battery will be prepared. Test administration materials will be prepared, and personnel designated as on-site test administrators will be trained by Task 2 staff.

14. Monitor administration of Experimental Battery to FY86/87 cohort, further analyses of FY83/84 cohort data. The administration of the Experimental Battery will be carried out by Army personnel on site at training schools. Task 2 staff will make several scheduled inspection visits as well as any unscheduled visits necessary to respond to problems. During this time, further analyses of the FY83/84 cohort data will be carried out.

15. Analyze FY86/87 cohort data and prepare final reports. Predictor response distributions, covariances, etc., of the FY83/84 cohort and FY86/87 cohort will be compared to ascertain if substantial differences occur because of attrition (in the concurrent FY83/84 cohort sample) and other factors. Relationships of the predictors to training performance will be analyzed. Draft and final instrument and technical reports will be prepared.

We turn now to a short explication of the relationships between subtasks. As Figure 2.2 shows, there are five major phases in Task 2. The first phase is the literature search and planning (Subtask 1) which provides information and direction for the next phases. The next two phases are: the selection, administration, and analysis of a preliminary battery of predictors (Subtasks 2, 3, and 9); and the development, pilot test, and analysis of a set of new predictors, called the Trial Battery (Subtasks 4, 5, 6, 7, and 8). Assuming approval by ARI, the Preliminary Battery will be a set of well-established "off-the-shelf" measures that best covers the relevant predictor domains as indicated by the literature search. The new predictors will be newly developed or modified measures intended to measure those predictor constructs deemed most likely to be effective for predicting soldier performance, as determined by the literature search and two rigorous, formal evaluation steps (Subtasks 4 and 5). The inclusion of the Preliminary Battery of well-established predictors at this stage in the project allows an early examination of the covariance of different types of predictors (than those currently used) with present pre-induction predictors, as well as an examination of the ability of these different kinds of predictors to predict performance during training. It also makes possible a predictive validity study of the measures in the Preliminary Battery since job performance information will later be collected (in Subtask 11 and in Task 5) on members of the FY83/84 cohort tested at this time.

Note that these two phases proceed somewhat in parallel. Early analyses of the Preliminary Battery (Subtask 9) will inform the development of the Trial Battery (Subtask 6). Thus, if these early analyses show that some

of the constructs measured in the Preliminary Battery are highly redundant with ASVAB or other current pre-induction measures or with other measures in the Preliminary Battery, work will not proceed on developing new measures of those constructs. Work on Subtask 6 will not be held up, however, since there will be constructs not included in the Preliminary Battery on which development work can proceed, particularly in the perceptual/psychomotor area.

The fourth phase (Subtasks 10, 11, and 12) includes the development and administration of the revised Trial Battery, an improved version of the new predictors, by building on inputs from Subtasks 8 and 9. In addition, Tasks 4 and 5 provide Army-wide and MOS specific job performance measures which make possible two validity investigations in this phase: a concurrent validity effort for the Trial Battery and a predictive validity effort for the Preliminary Battery (using as subjects those soldiers who completed the Preliminary Battery in Subtask 3 and for whom job performance criteria are collected in Subtask 11).

The final phase of Task 2, consisting of Subtasks 13, 14, and 15, includes the preparation of the Experimental Battery based on input from Subtask 12, administration of the Experimental Battery to members of the FY86/87 cohort, analysis of the relationship of the battery to training criteria and subsequent final revisions to the battery.

Although Task 2 formally ends at this point, job performance criterion data will later be collected by other Project A staff (from Task 5) and a

predictive validity analysis of the Experimental Battery will be completed by Task 1 staff.

We turn now to a delineation of the inter-task dependencies between Task 2 and Tasks 1, 3, 4, and 5.

Task 1. Task 2 and Task 1 will coordinate their work very closely. This is necessary because, in general, Task 2 staff will develop predictor measures and collect the data in the field, while Task 1 staff will receive the collected data and prepare and edit the data. With regard to analyses, Task 2 staff will perform those analyses directly related to development and refinement of measures, while Task 1 staff will perform the analyses directly related to validation of predictors. In reality, however, development of measures and collection of data have direct bearing on preparation and editing of data files. Likewise, analyses aimed at development and refinement of measures have implications for validation analyses. Finally, the results of validation analyses feed directly into refinement of predictors. This simply means that Task 2 staff will be responsible for close communication with Task 1 staff with regard to formats and content of predictor measures, methods of data collection, and anticipated developmental analyses. Task 1 staff will be responsible for providing guidance and advice on these matters, particularly with regard to anticipated problems in data preparation or editing and alternative methods of developmental analyses. Also, they will be responsible for communicating planned validation analyses to Task 2 staff, so they, in turn, may provide feedback

on the usefulness of those analyses for further refinement and development of measures.

Finally, all training or job performance criterion measures that are input to analyses performed by Task 2 or which impact on Task 2 efforts will only be available through Task 1, since they are managing the data base. On the other hand, conceptual information or constructs underlying training and job performance criteria will pass directly from Tasks 3, 4, and 5 staff to Task 2 staff.

Task 3. Task 3 staff will develop measures of training performance and be responsible for collecting these data. These data will be added to the longitudinal research data base (LRDB) by Task 1 and will then be available for analyses by Task 2. Subtask 9 and Subtask 15 of Task 2 rely on Task 3 providing training measures (through Task 1). In addition, Task 3 staff will provide to Task 2 staff conceptual information about training performance constructs, based on their early work with training school instructors and review of training measures. Task 2 staff will use this information in Subtask 4 (technical review), in September, 1983.

Tasks 4 and Task 5. Both of these tasks are responsible for developing improved measures of job performance and collecting these data. There are three major types of interaction between these two tasks and Task 2. The first kind of interaction is the provision to Task 2 staff of conceptual information about job performance. This information is input to Subtasks 2

(selection of a Preliminary Battery) and 4 (technical review of prediction constructs and measures). Staff from Tasks 4 and 5 have already provided information to Task 2 staff and that information is being used in Subtask 2. Task 2 staff has provided Tasks 4 and 5 staff with further guidance on the nature and format of their information requirements for Subtask 4, based in part on the job performance information already provided to Task 2. The information from Tasks 4 and 5 is required by September 1, 1983.

The second interaction is the provision of data on job performance. Task 5 staff will collect this information and provide it to Task 1 staff, who will make it available to Task 2 staff through the LRDB. These data are required for the analyses carried out by Task 2 staff in Subtask 12, beginning in October, 1985.

Finally, the staff of Task 5 is responsible for the collection of data on the predictor measures administered to the FY83/84 cohort during June-September, 1985. Task 2 staff will provide the predictor measures, administration manuals, and training required to administer the Trial Battery. The data collection teams from Task 5 will then administer the measures, with on-call assistance from Task 2 staff.

PROCEDURE

Subtask 1: Literature Search and Planning

Rationale. As mentioned earlier, the present set of pre-induction predictors does not include measures of several domains of human attributes that have been useful in other settings for predicting work performance and in classifying persons according to occupations for which they are most qualified. In this regard, there appear to be dozens of constructs in the perceptual, psychomotor, biographical and vocational interest, and cognitive domains that are not presently measured during current Army pre-induction screening procedures.

A listing of from 70 to 90 variables covering the major human attribute domains can be thought of as the "whole person" approach for identifying a relevant set of job performance predictors. In theory, if all major domains of human attributes were to be adequately covered by appropriate measures, then one would expect to be able to predict performance in almost any job. Even if this were to be the case, it is obvious that we will not have the luxury of measuring so many human attributes.

Thus, it is necessary in our research to narrow our gaze eventually to those constructs that have the greatest likelihood of meeting our aforestated emphasis on performance relatedness. The literature search, therefore, must identify as many potentially useful predictors as possible and provide information about those predictors in a manner useful for

Subtasks 2 and 4, selection of a preliminary battery and technical review of predictors.

In addition, this subtask i· ˋ·ˊes ˏubstantial effort devoted to planning. As the project began, new in.ormation has required some changes in project design and, consequently, revisions to Task 2 procedures and resource allocations became necessary. We have no doubt that constant attention to plans and, when necessary, modification of those plans, will continue throughout the project.

<u>Procedures</u>. For convenience, the potential predictor domain has been divided into three areas: (a) cognitive/perceptual; (b) vocational interest, and biographical; and (c) psychomotor abilities. A team has been formed for each domain consisting of a leader with one to three research associates or assistants, and one or more expert consultants. Each team has been searching and reviewing the literature within its domain. Team leaders report to the task leader who has coordinated the search.

1. <u>Review Forms</u>. Two forms have been developed to record information from documents as they are reviewed. The intent of these review forms is to capture the information so that a critical, technical review can be performed later, in Subtasks 2 and 4. (The initial reviewers will also make a critical review, but they primarily insure that adequate information is recorded for the later technical review.) The content of the review forms is :dentical across domain teams and includes information relevant to evaluating each potential predictor.

2. Search plans. Plans for searching each domain have been formulated. Predictors used currently or in the past by the Army and other services are being reviewed, as well as predictors used in public and private sectors. Published and unpublished literature is being searched. Appropriate computer searches of educational, psychological, business and government areas have been completed. Journals known to be highly relevant have been systematically reviewed. Domain experts have identified unpublished or "in press" research as well as researchers to be directly queried.

3. Management of search. Team leaders will monitor the completed reviews for completeness and accuracy. A list of citations will be compiled within domains and reviewed by team leaders and the expert to insure comprehensiveness of the review. Team leaders will organize the reviews by predictor content or construct and, if possible, cross reference by criteria "predicted." The task leader will review citation lists and a sample of completed reviews and meet with task leaders at least biweekly.

The task leader will also maintain contact with leaders of other tasks, especially Tasks 4 and 5, in order to obtain and update performance criteria information. As this information becomes available, its implications for the predictor literature search will be evaluated and used to redirect the search, if necessary. Thus, if delinquency (AWOL, drunk driving, drug abuse, etc.) appears to be an important criterion set, then research attempting to predict such phenomena will be reviewed, if it has not been reviewed previously.

4. Planning. The task leader has been (and will continue) to keep in close contact with his counterpart at ARI and with the project director, principal scientist, director of technical planning, and other task leaders of Project A. He has attended a course on PAC III in order to facilitate future resource planning. Written revisions of the research plan will be prepared as necessary, but telephonic, WYLBUR, and personal meetings among the persons mentioned above will be the primary method of keeping plans on track.

Subtask 2: Selection of Preliminary Battery and Preparation for Administration to FY83/84 Longitudinal Sample

Rationale. After the project began, it became apparent that it would be desirable to collect data from a relatively large sample of soldiers on new pre-induction predictors at a point somewhat earlier in the project than originally planned. With regard to Task 2, there are two primary reasons for administering a preliminary battery. First, the collection of data on a number of new predictors that comprehensively represents the types of predictors not currently in use will allow an early determination of the extent to which such predictors contribute unique variance, or actually measure human attributes not measured by current pre-induction predictors. This information will be useful for guiding the development of new predictors into areas most likely to be useful for increasing the accuracy of prediction and classification. Second, the early collection of preliminary battery data on soldiers during their advanced training phase allows the conduct of a predictive validity investigation using new pre-induction

predictors much earlier in the project. Thus, empirical data on the predictive validity of new predictor constructs will be available 36 months after the project begins.

The purpose of this subtask then is to make a careful selection of measures of new pre-induction predictor constructs in order that advantage is taken of the design features just described.

Procedures. The literature review forms will serve as the primary input to a careful critical review of potential measures for inclusion in the Preliminary Battery. The Preliminary Battery must necessarily be made up of "off-the-shelf" instruments because there is too little time prior to the scheduled administration of the Preliminary Battery to develop and pilot test new measures of constructs deemed potentially useful. This means, for example that a single published interest instrument that is judged to best cover the constructs in that area will probably be chosen, rather than selecting several scales from each of several different instruments and printing a new instrument containing those scales.

Assuming approval (for research purposes) of the use of "off-the-shelf" measures by ARI, the review process for selection of the battery will consist of two major steps: an internal review by Task 2 staff with cooperation by the ARI Task 2 monitor and a presentation to and subsequent review by ARI of the candidate predictors selected for the Preliminary Battery. The internal review will proceed as follows: each domain team

will prepare an initial list of possible predictor measures, organized within fairly broad constructs (List 1). The teams will then apply the 13 technical predictor evaluation criteria (previously submitted to ARI) to narrow this list to those predictors that are serious possibilities. Thus, predictors with major problems with regard to one or more of the 13 evaluation criteria will be eliminated. In addition, those predictors that are not readily available for large-scale administration, i.e., are not really "off-the-shelf", will be eliminated (List 2). List 2 will then be examined with regard to the 13 criteria, but in a comparative sense. Thus, predictors will be compared to each other, as well as against the evaluation criteria. Also, the job performance criterion information presently in hand from Tasks 4 and 5 will be cast against the candidate predictors. This process will result in a third list that contains the recommended predictor measures and one or more alternatives, all ranked by preference. List 3 will be presented to the task leader and all other Task 2 staff in a review meeting and the choices of predictors will be examined and modified, if necessary, during this meeting. We think it would be extremely useful for the ARI Task 2 monitor to participate in this review meeting, which will occur during the last part of April or first week in May.

Following this meeting, the recommended Preliminary Battery, including suitable alternative measures, will be presented to ARI by the Task 2 leader. This could best be accomplished in a one-day meeting in Washington in mid-May, 1983. After approval of the battery by ARI and other necessary reviewing authorities, Task 2 staff will procure materials necessary to administer the battery, make arrangements with publishers, and carry out

any other activities necessary to secure the instruments in sufficient numbers to complete the administration of the battery.

The next major step will involve planning the administration process and preparing manuals and training materials for use by the on-site administrators. As part of the preparation, we will pretest the selected instruments on a sample of 40 soldiers in training or in a variety of MOS, from either TRADOC or FORSCOM. The purpose of this pretest is to identify problems with test instructions and logistics of the administration process. Four hours of soldier time would be required. This pretest would occur in July, 1983.

Two points need to be raised here about the Preliminary Battery. First, no computer-administered measures will be included due to the short lead time available to prepare for the administration. (We note, however, that such measures will be intensively investigated and developed as part of the development of the Trial Battery.) Second, initial inquiries of TRADOC indicate that soldier time at AIT schools is generally allocated in four hour blocks. We think four hours is sufficient time to administer "off-the-shelf" measures of biographical information, vocational interest, and cognitive/perceptual ability. These measures can be group administered. Physical and psychomotor measures present some difficulty, however. Almost all of these measures require individual administration. It appears to us at this time that the administration of a sufficiently comprehensive set of individually administered psychomotor tests <u>and</u> the group administered

measures will be very difficult to accomplish in four hours. Furthermore, the administration of individually administered measures requires more expertise on the part of the administrator than may be available on site. If this is the case, much more training of administrators would be required.

These problems could be overcome by obtaining more soldier time and expending more resources on training admininstrators or by electing to cut back somewhat on the range of constructs covered in the Preliminary Battery. We favor the latter option at this time. Recall that one of the primary reasons for the Preliminary Battery is to determine the amount of unique variance that would be contributed by new predictors. The psychomotor measures are perhaps the least likely of the potential set of new predictors to correlate highly with the major current predictor, the ASVAB. Therefore, there seems much less cause for concern if they are not included in the Preliminary Battery. Again, we point out that psychomotor measures will be investigated as part of the Trial Battery development. Measures of bio-data, vocational interests, and cognitive/perceptual tests are of much more interest with regard to their overlap with the ASVAB and with each other and should definitely be included.

As part of this subtask, a detailed outline of the literature review report will be prepared and delivered to ARI. (The full report will be prepared

and delivered after the Preliminary Battery is in the field, i.e., in Subtask 4).

Subtask 3: Administration of Preliminary Battery to FY83/84 Longitudinal Sample

Rationale. The preliminary battery will be administered to soldiers in AIT for four MOS: 05C (Fort Gordon), 71L (Fort Jackson), 63B (Fort Dix and Fort Leonard Wood), and 19E/K (Ft. Knox). These four MOS were picked to represent a diversity of job types and because they had sufficient numbers going through AIT to meet sample size requirements, i.e., enough tested soldiers in these MOS will still be in the Army and available for collection of job performance criterion data in June-September, 1985 at the sites visited (see INTRODUCTION section on sampling). Local on-site administrators will be required to administer the preliminary battery. This is necessary because it is impractical for Task 2 staff to travel to the five sites each time a new class begins (the soldiers will be tested during the first week of training, see Procedures for this subtask) or to live on-site for the long time period over which the data will be collected (October, 1983-June, 1984). Thus, Task 2 will be responsible for training these administrators and monitoring the administration process.

Procedures. During the month preceding the beginning of data collection (September, 1983), Task 2 staff will visit each of the five administration

sites for approximately one week. Prior to these visits, arrangements for selecting the on-site administrators will be completed. The administrators will in all likelihood be either active duty Army personnel with appropriate backgrounds in training or personnel, i.e., general familiarity with testing, training, or personnel work; or local, contracted personnel with appropriate experience. We have been and are now making inquiries at TRADOC posts about the availability of personnel for test administration. These inquiries, to date, indicate that active duty Army personnel with appropriate experience are present at the sites (primarily, these personnel are AIT instructors) and could administer the tests. We do not, of course, have committments for such personnel since that can only come through appropriate troop support requests. Use of Army personnel, rather than local contracted personnel, would avoid incurring expenses not originally budgeted.

Regarding the selection and training of test administrators, we prefer to rely primarily on training. Task 2 staff have had extensive experience in preparing test administration procedures and manuals and in providing training for test administrators. We have trained persons with little experience in testing in several large-scale validation studies and achieved satisfactory results in terms of quality of data (Dunnette, et al., 1981; Peterson & Houston, 1980; Peterson, Houston, & Rosse, in press). A minimum of two administrators should be selected to be trained. This number depends on the size of classes and desired size of the pool of trained administrators. It could be as large as 10 to 12 persons at some posts. It is essential that one person at each site be designated as the primary contact. This person does not necessarily need to be involved in

the actual administration of the battery but will be responsible for securely storing all supplies, shipping completed batteries to Task 2 staff, monitoring attendance at testing sessions, and communicating with Task 2 staff.

The site visit activities will include a one-day training session for the administrators on administration procedures, inspection of facilities, completing scheduling of sessions and other details. At this point, we believe the facilities required will be one or two air-conditioned classrooms large enough to accommodate at least 50 persons, equipped with student desks. (One classroom will normally be sufficient for some sites, based on TRADOC estimates of class size, but some classes at some sites are large enough to require two or even more classrooms in order to complete testing of an entire class during the same day.)

After the initial site visits, Task 2 will communicate by telephone at least weekly to monitor administration activities. In addition, completed batteries will be sent to Task 2 staff as soon as they are completed and these will be inspected to detect any abnormalities or problems. As problems surface, Task 2 staff will return to the sites to assist in their solution. Even if no problems surface, at least two "monitoring" site visits will be made to observe the administration procedures.

Two final comments about data collections are in order. First, the Preliminary Battery should be administered to soldiers during their first week at AIT, if possible. Recall that the Preliminary Battery scores will be

correlated with training performance (Subtask 9) and later, with job performance. Thus, the earlier in AIT that the battery is administered the less opportunity there is for the training itself to influence scores on the battery and, therefore, contaminate the correlational analyses.

Second, we have planned our schedule to begin testing in October, 1983. There are two reasons for this:

1. The summer months see a large influx of National Guard and Reserve soldiers for training. This complicates the administration process in the sense that we cannot use these soldiers for purposes of the predictive validity study, and TRADOC has advised us that it is administratively very difficult to separate regular soldiers from Reserve and National Guard soldiers within classes for purposes like the Preliminary Battery testing. That is, each class is treated as a single unit and apparently 40-60 percent of a class could be Reserve or National Guard during the summer.

2. There are at least two other programs requiring testing of soldiers at TRADOC for some period during the summer--the Basic Skills Educational Program (BSEP) and validation work on a physical fitness battery.

These two facts make an October start date more feasible than a summer start date. We have obtained estimates of the input for the four MOS and an October start date would still permit the collection of the targeted sample sizes in two of the four MOS (and nearly so for the other two).

Draft field test plans containing full details on tests to be administered and facilities required will also be prepared as part of this subtask, after the content of the Preliminary Battery has been determined. A final field test plan will be prepared after ARI review.

Subtask 4: Technical Review of Predictor Constructs and Measures

Rationale. A great deal of information will be discovered and reviewed during the literature search. This information must be subjected to a careful, thorough review in order to identify the "best bet" set of predictor constructs and measures. One part of this subtask is designed to achieve that goal by using a formal judgment process employing experts. The method for this subtask has been used successfully by Bownas and Heckman (1976); Peterson, Houston, Bosshardt, and Dunnette (1977); Peterson and Houston (1980); and Peterson, Houston, and Rosse (in press), in identifying predictors for the jobs of firefighter, correctional officer, and entry-level occupations (clerical and technical), respectively, and Peterson and Bownas (1982) provide a complete description of the methodology. In this technique, descriptive information about a set of predictors and the job performance criterion variables are given to "experts." The experts make estimates of the relationship between predictor and criterion variables, generally with a five-point scale or even by directly estimating the correlation coefficients. The final result is a matrix with predictor and criterion variables as the columns and rows, respectively. Cell entries are experts' estimates of the degree of relationship between the particular predictors and various criteria. The interrater reliability

of the experts' estimates is first checked. In general, these reliabilities have been impressive, in the .80 to .90 range for about 10 to 12 experts. If the estimates are reliable, then the matrix of predictor-criterion relationships can be analyzed and used in a number of ways. The covariances of the predictors can be estimated by using the profiles of their estimated relationships with the criteria, that is correlate the columns. These covariances can then be factor analyzed in order to identify predictors that function similarly with regard to predicting the job performance criteria. The covariances of the criteria can be similarly examined, and criteria likely to be predicted by a common set of predictors can be identified. In addition, equations relating predictors to criteria can be derived for either the predictor variables or factors. In this way, redundancies and overlap in the predictor set can be identified and an efficient, integrated set of predictors chosen to carry forward into later phases of the research.

Previous use of this methodology (studies cited above) shows clearly that when content-valid performance criteria, i.e., based on accurate job information derived from task analyses, critical incident studies, etc., and predictor measures found through a careful literature search are used by experienced psychologists and other experts, reliable estimates are obtained. Moreover, predictors selected and weighted according to such estimates have shown significant, empirical relationships with measures of job performance in follow-up studies.

In addition to the formal judgment process described just above, another type of technical review will be carried out for the perceptual/psychomotor

domain of measures. In the course of our literature search and initial review of research in this area, it has become apparent that relatively little research, especially criterion-related validation work, has been completed on recently developed computer-administered perceptual/psychomotor measures. This method of measuring perceptual/psychomotor measures, however, has great practical and theoretical advantages over the older kinds of methods using more cumbersome mechanical and electrical apparatus. It is our present judgment that perceptual/psychomotor measures must only be pursued via computer administration. It makes little theoretical or practical sense to attempt to validate and operationalize the 1950s technology of measuring such variables. Therefore, we will carry out a less formal and more wide-ranging review of predictor measures in this area that will take place prior to the formal, expert judgment review. It is desirable to carry out this review as early as possible so that we may quickly begin developmental work on these measures (in Subtask 6).

Procedures. The first two activities in this subtask related to the formal judgment process are the development of the definitions of the rows (criterion constructs) and columns (predictor measures) of the judgment matrix. With regard to the rows or criterion constructs, Task 2 staff has requested information from the staff of Tasks 3, 4, and 5. The product that Task 2 staff will receive from the three other tasks is similar in format, to wit: the name of the criterion construct or dimension; a brief definition of the construct; elaborations, examples, illustrative materials, or other further explanatory information; a brief description of the data base, analytic methods used, or other information that would allow Task 2 staff to properly evaluate the information. The substantive

content will, of course, vary across the three other tasks as will the sources of information they draw upon to develop their set of criterion constructs. After receipt of these lists of criterion constructs from Tasks 3, 4, and 5, Task 2 staff will review, edit, and integrate them as required for the technical review process. (Task leaders from Tasks 3, 4, and 5 have informed the Task 2 leader that they can develop such lists and deliver them to Task 2 staff by September 1, 1983.)

With regard to the columns or predictors, the task leader and team leaders will critically evaluate all predictor information contained in the literature review forms completed during the literature search (Subtask 1). Sets of "best bet" predictors will be chosen based on application of the evaluation criteria mentioned in Subtask 2. If sufficient information is available for predictors, Bayesian priors will be computed as outlined by Schmidt and colleagues (Schmidt, Hunter, & Caplan, 1981). At this point, more predictors than would be feasible to study further will be retained. Within each domain, the selected predictors will be placed in one of three categories: (a) predictors with existing, adequate measures; (b) predictors with existing measures, but requiring some modification prior to Army use; and (c) predictor constructs or variables with no existing measures or measures that require almost total development. (Only predictor constructs with very strong theoretical or content-related promise would be retained in the third category. On the other hand, such "new" predictors might very well be the most desirable in terms of predicting portions of the criterion space not now predicted.) Task 2 staff will then prepare a packet for each predictor that provides a

concise, comprehensive description of the construct measured, reliability, validities, adverse impact, etc.

Then, an appropriate set of experts will be asked to complete the matrix, as described in the rationale section of this subtask. The experts should be industrial, measurement, or differential psychologists with experience and knowledge in personnel selection research and/or application.

Some 8 to 13 of the experts will be drawn from the PDRI, HumRRO, and AIR researchers working on this task: the task leader (1), leaders of the various predictor domain literature searches (2-3), the consultant experts for the predictor domains (3-5), and several of the research associates who have worked on the literature search (2-4). ARI psychologists will also be asked to complete the task.

The ratings can be completed without a meeting of the raters. Telephone communication is generally sufficient. Completing the ratings should require one to two days.

The ratings will then be analyzed by Task 2 staff as described in the rationale section. The data will be submitted to Task 1 staff for inclusion in the data base. We note here that the analysis of the expert ratings will be the primary mode of selecting the technically best set of predictors for further development. However, we will carefully examine the possibility of including some measures that were included in the Preliminary Battery in the Trial Battery, even if these measures do not survive technical review. Such inclusion is desirable in order to have

well-defined, reliable marker variables for factors analyses of the Trial Battery, which will occur in Subtask 8. In addition, comparisons of predictive vs concurrent validity can be made with these common predictors when they are administered twice to the FY83/84 cohort.

The technical review for computer-administered, perceptual/psychomotor measures will have three steps. First, the relevant predictor review forms, i.e., those describing perceptual/psychomotor measures, from the literature search will be critically evaluated just as in the formal judgment process. Second, Task 2 staff will make visits to Army and other services locations that are currently engaged in developing or validating such measures. These site visits will include at a minimum, Fort Rucker, Fort Knox, and the Air Force Laboratory in San Antonio. The general purpose of these visits is to learn as much as possible about current development of these measures. More specifically, we will focus on hardware configurations, available software and problems in software development, psychometric properties of these types of measures, especially with regard to reliability and practice effects, criterion-related validity data, and information on the kinds of job performance constructs that are being predicted and/or simulated. After the completion of these two activities, Task 2 staff will review the available information and identify the perceptual/psychomotor constructs that appear to be most worthy of further development. This review will use the 13 evaluation criteria used for all other predictor measures, plus other criteria that may emerge as a result of the first two steps. These criteria will be applied against written descriptions of the perceptual/psychomotor computer measures, derived from

the first two activities in this review process. Furthermore, we will inquire, during the above-mentioned site visits, about the availability of the researchers at those sites for participation in this review process. Several members of Task 1 staff possess expertise in this area and they will participate in this review step as well.

Finally, a literature review will be prepared as part of this subtask. It will summarize the major findings from our literature search and review, and will be prepared in accordance with the outline submitted to ARI as part of Subtask 2.

Subtask 5: Cost/Administrative Practicality Review

Rationale. At this point, we will have a set of predictor constructs that, collectively, are the best possible from a technical point of view. To prevent wasting time and money on validating predictors that cannot ultimately be operationally administered, a cost/administrative practicality review must be undertaken.

Procedures. Information about the set of predictors that passed technical review will be prepared. The information will include a definition of the variable or construct; content description or item examples; method, time required, and costs of administration; time, costs, and nature of development efforts; and any other information necessary for a review of the practicality of a predictor. No psychometric or technical information will be included, since the predictors will already have passed technical review.

A cost/administrative practicality review panel will be selected. Members should have, collectively if not individually, knowledge of the field operation of recruitment, selection, and classification of soldiers, privacy concerns, human subjects review policies, and other administrative or practical issues relevant to pre-induction testing. Preliminary inquiries indicate that appropriate panel members can be identified and are willing to serve on such a panel. Some members of this panel should come from the support officers that will handle troop requests. In this way, the military personnel assisting in securing troop requests will have an early familiarity with the type of predictors to be field-tested in later subtask activities.

Panel members will receive the predictor information about two weeks prior to a two- to three-day workshop, along with instructions and forms for recording their initial judgments about the predictors. At the workshop, the task leader and team leaders will work with the panel members to retain or delete predictors based on estimates of cost and praticality. An attempt will be made to strike a balance between the estimated predictive effectiveness and psychometric adequacy of a predictor and the estimated practicality and cost of a predictor's development and operational administration.

Predictors passing this review will move forward to the next subtask. Note, however, the predictors not passing this review are predictors that possess technical merit but are presently too costly to develop further in

Project A. Conditions that led to these cost decisions are certainly subject to change. Furthermore, some of these predictors could be developed in other research efforts. The point is that these "rejected" predictors actually constitute a separate research agenda.

Subtask 6: Initial Development of Predictors for the Trial Battery (New Predictors)

Rationale. The predictors surviving technical and cost reviews will require varying degrees of further development. The purpose of this subtask is to complete the initial developmental steps, try out the predictors on small samples of soldiers, and revise the predictors. As described in Procedures for this subtask, we think an iterative process of several tryouts and revisions will be the most efficient way to complete this work.

Procedures. Two different types of development efforts will be required. One effort will involve the writing and revising of paper-and-pencil cognitive and non-cognitive measures. The second will involve the development of computer-administered versions of perceptual/psychomotor measures and computer-administered versions of some of the paper-and-pencil measures. We plan to begin efforts on the computer-administered perceptual/psychomotor measures immediately after the technical review process for those types of measures, in July, 1983. Development of paper-and-pencil measures will begin later, after the cost/administrative practicality review or December, 1983. The early start on computer-administered measures is desirable because of the greater amount of development work required. It is also possible that a third type of developmental effort will be

required. This effort would focus on the development of tests of psychomotor abilities using apparatus that does not require computers. As we have already stated, we do not think it advisable to test psychomotor abilities without computer administration. Thus, we believe it is unlikely that this third developmental effort will be required. If it does turn out to be required, however, this effort would begin at approximately the same time as the paper-and-pencil measure development begins, and will follow the same timetable for tryouts and revisions as that effort.

The sequence of activities in the development of computerized, perceptual/psychomotor measures will be as follows:

1. Identifying and obtaining the appropriate hardware for initial development, as determined by the technical review. A minimum amount of hardware will be obtained for these initial efforts.

2. Writing software or modifying software obtained from other development efforts (identified in the field visits that are part of the technical review in this area) to measure the constructs identified in the technical review.

3. Trying out the measures on a small sample of soldiers (N=10). This will occur in November, 1983, and the try out could take place at a MEPS site. The focus is on debugging the measures, obtaining

feedback from soldiers on the acceptability of the measures, and getting a reading on the administrative problems involved. The only facilities required will be a classroom with furniture appropriate for setting up the computer hardware (tables and chairs) and normal household electrical supplies.

4. Revising the software and/or hardware in light of the initial try out results.

5. Conducting a second try out on a slightly larger sample (N=30). This will occur in January, 1984, and it would be preferable to try the measures at a MEPS site.

6. Writing a preliminary report on the results of the initial development and try outs. The purpose of this paper is to advise ARI of results to date and provide a judgment about further developmental work and associated hardware costs. This report will be submitted by March 1, 1984, in order to provide ARI and the contractors time to make a judgment about the costs and administrative feasibility of including these measures in the pilot test of the Trial Battery (Subtask 7).

7. Should the decision be to go ahead with these measures, then further software development will be carried out and arrangements for the necessary hardware to carry out the pilot test in Subtask 7 would begin.

8. A final try out of the measures would be tried out in May, 1984, on a sample of forty MEPS candidates, concurrent with the try out of the paper-and-pencil measures that may be objectionable in some way and the computer-administered versions of some of the paper-and-pencil measures. (Reception Stations could substitute for MEPS in 3, 5, and 8).

The development of computer-administered versions of some of the paper-and-pencil measures would begin after the paper-and-pencil versions were initially developed, approximately February, 1984 and these measures would be tried out during the April, 1984 try out of the paper-and-pencil measures (see below). Although these computerized measures are conceptually distinct from the computerized psychomotor/perceptual measures, we think their inclusion in the pilot test is linked with the more important decisions about the scope, nature, and expected payoff of using the psychomotor/perceptual measures in the pilot test, and in a major sense is dependent on those decisions.

Finally, we intend to keep ourselves fully informed on the developments in the Army and other services with regard to computer-assisted testing hardware and software. These developments will be very important to the development efforts just outlined here and, especially, to the ultimate feasibility and practicality of any computerized measures under development.

The activities devoted to the initial development of paper-and-pencil measures of non-cognitive and cognitive abilities will be as follows:

1. Preparation of specifications for measures. We will have accumulated a great deal of information that will be very helpful for writing or modifying items to be included in these measures. The literature search and technical review will provide the basic information. This information will be organized by the Task 2 leader and domain leaders into a package for each measure that will contain a concrete, specific definition of the construct to be measured, examples of existing measures of the construct, specification of the item format and response format, desired number of items, and a description of the administration procedures that are desired.

2. Task 2 staff will then write items for the measures and these will be reviewed by the domain leaders and the task leader and corrective feedback will be given. Our expert consultants will review these measures as they are developed.

3. The first try out of the paper-and-pencil measures will occur in March, 1984, about 14 weeks after development begins. A total of 60 soldiers (from FORSCOM or TRADOC) will be requested for one day. There are no special MOS requirements, except that the soldiers should represent a diversity of jobs. To the extent possible, both sexes and the major race groups should be represented. As noted below, however, no significant statistical comparisons of group performance will be carried out on data collected during the try out. The intention here is to avoid trying out the measures on a homogeneous group of soldiers, instead obtaining data from a heterogeneous group. These soldiers will be broken into two groups of 30 soldiers each. They

will attend two sessions, one in the morning and one in the afternoon. Soldiers will complete different sets of predictors at the two sessions. After each session, they will complete short evaluation forms about the predictors and will be asked about their reactions to the measures.

The focus of these try outs is on "debugging," i.e., the efficiency, practicality, and understandability of test items, instructions, and administration procedures. Only very simple statistical analysis, e.g., simple frequency contents of items completed, will be conducted on these data so there is no need for larger sample sizes. The target group size of 30 is based on our past experience with developmental efforts. This size of group provides sufficient diversity, but is not so large as to inhibit the elicitation of direct feedback from participants that is necessary at this stage of development.

Two normally equipped classrooms will be required to conduct the try outs, i.e., with student desks or tables and chairs.

4. Information from this try out will be used to revise the measures. A second try out will be conducted approximately one month after the first try out. This try out will be conducted similarly to the first one. Measures will again be revised.

5. A third try out of some of the paper-and-pencil measures on applicants at a MEPS station should occur approximately one month after the second try out. The purpose of this try out is to get an early

reading of the reactions of candidates for Army EP ranks to those measures that may be objectionable in some way (some biographical or personality measures). As we stated above, this sample will also complete the computerized measures. We think a sample of 40 candidates for four hours should be sufficient. If possible, the MEPS station selected should be typical, in the sense that it should not have an atypical population of candidates being processed. The 40 candidates should be heterogeneous with regard to race and sex background. No major statistical analyses of their data will be conducted just as in the other tryouts.

All measures, computerized and paper-and-pencil, will be given final revisions based on this last tryout, in preparation for Subtask 7, the pilot test.

Subtask 7: Pilot Tests of Predictors for the Trial Battery (New Predictors) in the Field

Rationale. The pilot testing subtask is designed to answer three important questions: (1) Do the newly developed predictor measures work administratively for fairly large samples? (And, related to this, how are the measures received by soldiers?) (2) What are the item and test characteristics of the new measures, i.e., item response frequencies; test, scale, and item reliabilities; stabilities or test-retest reliability test score distributions? (3) How do the new tests covary with each other, with current pre-induction measures, and available criterion information?

In addition, fakeability of responses will be of some concern for bio data and vocational interest measures. Research addressing this concern is appropriate for this stage of the effort. One sample of soldiers will be asked to answer these items to make themselves "look as good (or well-qualified) as possible" and another to "look as bad (or least-qualified) as possible." A possible variation on the instructional set will be to ask soldiers to "answer so that you would be chosen for a job in the electronics field" or some other target field.

A second examination of fakeability will require administration of the bio/vocational interest predictors to a group or groups of candidates at MEPS sites. These persons will presumably be interested in doing as well as possible, even though they will, of course, need to be debriefed after taking the battery so that they understand that it had no impact on actual selection or classification decisions for them.

One other special effort will be required. Practice effects, beyond the normal test-retest phenomena, may be of concern for some of the perceptual/psychomotor tests. Therefore, several test opportunities in quick succession should be given to a sample of subjects. It is possible that practice is a necessary part of the administration protocol for tests of this type; thus, it will be necessary to evaluate such practice effects to determine when the most stable between-subjects measure of psychomotor abilities may be obtained.

Fakeability and practice effects are only two factors that can have undesirable (or unknown) effects on test scores. Others are order effects such

as fatigue, and context and situation effects such as lighting, administrator, amount of work space, etc. We have isolated practice and fakeability (or deception and malingering) as two important extraneous factors to be examined early in the research. In general, our approach to these extraneous factors will be to exert direct control whenever possible, and to conduct appropriate research to estimate the magnitude of such effects if direct control is not possible. Situational effects can usually be directly controlled by specifying the physical conditions of test administration and insuring adequate training for all administrators. Order effects can be estimated or controlled. That is, all subjects can be given tests in the same order which would be adequate unless there is an interaction between subjects and order. The only way to estimate such interaction effects or certain other effects such as deception or deliberate bias is to carry out research, similar to that outlined for fakeability and practice effects. (We believe the proper place for examination of order effects is later in the project when the final, smaller version of the new predictor battery is being put together. This later version will contain tests more nearly like those that would be operationalized and the order information would be most useful then).

All research on the effects of extraneous factors are concerned with the reliability, or more accurately, the generalizability of the measurement process (Cronbach, Gleser, Nada, & Rajaratnam, 1972). (This concept is <u>not</u> the same as the <u>generalization of validity</u> which is concerned with the degree to which a test's relationships with other variables are the same across extraneous factors; rather, this use refers to the extent that

persons' test scores will be the same across extraneous factors.) Ideally, all extraneous factors would have little or no effect on a person's test score, but these effects must be examined to determine if that is the case, and if it is not, to determine what procedures can be brought to bear to reduce the impact of such effects. (It is possible, also, as mentioned previously, that practice on performance measures is an important part of the administration protocol; that is, scores on later, more experienced trials may be better in predicting targeted criterion outcomes.)

We plan to employ the generalizability theory approach to the study of predictor reliability, as outlined in Cronbach, et al. (1972). This approach calls upon the researcher to clearly define the universe of generalizability, that is, to identify the facets that he wishes to generalize over (such as practice, order, administrator, or situation), and then to devise research designs to include those facets. ANOVA is then used to estimate variance components associated with each facet, enabling the investigator to determine the limits of generalizability of test scores and appropriate directions to proceed in order to increase the reliability or generalizability of the measurement process. Even if the actual ANOVA computations and variance component estimates are not completed, the analytic exercise of identifying all facets and thinking about their possible effects aid in defining samples and setting up test administration procedures.

The pilot test also affords the opportunity to look at the relationships of the experimental predictors to current pre-induction measures, training school performance, and job performance provided those scores are available

for the soldiers who participate in the pilot test. Use of soldiers from the FY81/82 cohort would provide the best opportunity for these analyses. These soldiers will have on file scores on their pre-induction measures and, perhaps, training school scores and some Army-wide measures. Task leaders from Tasks 1, 3, and 4 will be consulted about obtaining such data for the selected sample.

In addition to providing data necessary to refine the experimental predictor battery, these pilot tests will also provide a "shakedown" for the administrative procedures, coordination, and communication of the research teams, providing valuable information for the much larger cohort administration that will occur after revisions to the battery.

Procedures. Before detailing procedures at the pilot test sites, we will say a few words about sample sizes. There are essentially two types of concerns about sample size for the pilot test. First, we wish to have sufficient sample size to obtain stable estimates of covariance between predictors. For example, N=480 is sufficient to detect a correlation at .09 as being significantly different from zero at the .05 alpha level (Walker & Lev, 1953, p. 252) and provides a 95 percent confidence interval of ± .09 around Fisher's z transformation of r. For a sample value of .65, the confidence interval would cover the range of r's from .60 to .70. This degree of precision in estimating predictor covariances is sufficient for this stage of the research. Second, as we noted in the Rationale section for this subtask, we wish to carry out studies of fakeability and practice effects. In the main, the analysis of these effects will consist of tests of significance of mean differences between groups or simple ANOVA's.

Precise estimates of necessary sample sizes for such analyses cannot be made at this time using the usual power formulas (Walker & Lev, 1953) because some of the parameters needed to make the estimates are unknown. (The variances of the newly developed experimental tests cannot, of course, be even crudely estimated nor, likewise, can the size of practically significant score differences between groups on the experimental tests be estimated.) However, we can make realistic assumptions about these parameters now in order to make the best possible estimates of samples required. If we wish to detect mean score differences between groups of interest (say, a "regular" group and a "fake good" group) on the order of .25 standard deviations with an alpha (probability of Type I error) level of .05 and a beta (probability of Type II error) of .30, then we can compute the size of the required sample according to a derivation of the formulas given by Walker & Lev (1953, p. 166).

$$N = \left[(1.414)\ \sigma/d\ (Z_a + Z_b)\right]^2$$

where N = sample size

σ = standard deviation of the measure

d = size of score difference desired to detect

Z_a = Z score for specified level of alpha

Z_b = Z score for specified level of beta

Substituting the values for our case, the computation is:

$$N = \left[1.414\ (4.00)\ (1.645 + .253)\right]^2$$

$$N = 115.24$$

Therefore, sample sizes of 115 for the special studies of practice and fakeability will enable us to detect real differences as small as one-quarter standard deviation with a probability of .70 (1-beta) and with the alpha level at .05, i.e., the probability that we will decide there is a difference in scores between the two groups when there is no real difference. As the formula shows, decreases in level of alpha and beta or in the size of the score difference one wishes to detect will all result in increases in the required sample size. At this point, we think the levels stated above are adequate for pilot test purposes. With regard to the fakeability study, we will have developed a priori scales intended to detect faking and we think a sensitivity to mean differences on the order of .25 SD is sufficient for such scales as well as for scales measuring predictor constructs. The percentage of overlap between two distributions with a .25 SD mean difference is 90 percent (Dunnette, 1966, p. 143). (Samples of 115 would not be sufficient for doing empirical keying to detect faking, but as just stated, that is not intended at this point.) We are less certain about the sufficiency of an N of 115 for an investigation of practice effects, but are not now in possession of the necessary information to make a more precise estimate.

The implication of these estimates of sample size requirements for pilot test data collection are these:

1) An N of 480 for estimating correlations between predictors.

2) An N of 115 for the special studies of practice and fakeability.

We propose the following method of obtaining the pilot test data. The data would be collected at one post, if at all possible, on the FY81/82 cohort. One post is desired because in the event that physical/psychomotor test apparatus and micro computers will be required for administration, it will be extremely costly to have duplicate sets of these, and they are not easily transported. Furthermore, the time frame for pilot test administration does not appear to provide sufficient time to travel to several posts in succession. The FY81/82 cohort is desirable because these soldiers will be in their first tour but through training and will provide the opportunity to obtain scores on current pre-induction measures and in training for comparison and analysis with the experimental predictors.

The data collection method requires four distinct episodes of data collection. The first episode is the collection of the main body of data on a sample of 480 soldiers over a two-week period. As we now see it, each soldier will complete all predictor measures over a two-day period with the two days separated by one week. For purposes of this plan, we are assuming that earlier research (in Subtask 6) has indicated the desirability of pilot testing computer-administered measures and that sufficient hardware resources have been obtained. Furthermore, we are also assuming that some form of physical ability testing apparatus will be pilot tested. If either or both of these events do not occur, then the procedures outlined here will be much less complex and demand less in the way of soldier time. Given these assumptions, then, there will be three testing sessions:

1. a "paper-and-pencil" session in which the non-cognitive measures of vocational interests, bio-data and cognitive ability will be given. This session should require four hours, but may require six. Forty-eight soldiers would attend each session.

2. a "computer" session in which the computer administered measures of perceptual/psychomotor abilities and the alternative computerized measures of some of the paper-and-pencil battery will be given. Two groups of 12 soldiers would constitute a "session." At this time, we think two sessions of 24 per session can run each day, given the availability of twelve micro-processors.

3. an "apparatus" session in which the perceptual/psychomotor tests will be administered. The session is so named because there will be some apparatus involved as part of some of the tests. We plan to process 24 soldiers through each session in about three hours.

Each soldier will complete the "paper-and-pencil" session in one day, and the "computer" and "apparatus" sessions (one in the A.M. and one in the P.M.) on the second day. Using this method, 480 soldiers will complete all predictors over a two-week period with each soldier away from normal duties for 1-1/2 to 2 days. Ideally, these 480 soldiers will be made up of a variety of MOS.

The second episode will be the administration of the measures for which practice effects are a concern. This will primarily be the "apparatus" and "computer" sessions, though not all measures in these two sessions may be involved. This episode will require 115 soldiers not included in the original 480, and will require one day from each soldier. It will not be possible to take race and sex differences into account here, but it would be best to have an approximately equal number of males and females to make up this group. This episode will take place in the week following the first episode on the same post, for the reasons of practicality and economy already cited.

The third episode takes place at a MEPS site. One hundred and fifteen applicants will complete the parts of the paper-and-pencil battery that are of concern with regard to fakeability or motivational set (the non-cognitive measures). We will administer some of the "computer" measures in order to gauge their practicality for MEPS and accceptability to applicants. This group should also be equally split between males and females, if possible. (We are assuming no race or sex interaction with motivation. Sample sizes must be multiplied by at least four if this assumption is not made. At present we think this assumption is defensible.)

The fourth episode is the collection of test-retest and fakeability data. The test-retest will be collected, necessarily, at the same post as the first episode. The original sample of 480 will be asked to complete one of the three sessions a second time, or a half-day for each soldier. This yields a sample of N=160 to compute stability coefficients for each measure. The soldiers will be scheduled, if possible, so that race and sex

composition is balanced across the three sessions. An additional, separate sample of 230 soldiers will be required to complete the measures for which fakeability is a concern (the non-cognitive parts of the paper-and-pencil battery). One group will be instructed to "fake good" and one group to "fake bad." These groups should, if possible, be equally split between males and females. A half-day of each soldier's time will be required.

Task 2 staff will administer all measures, but we will require the assistance of one to three Army personnel to proctor sessions for the main data collection (episode one). Facilities required for the main data collection are three classrooms. One with student desks or tables and chairs for 50 persons, one cleared for placement of apparatus, and one classroom with tables and chairs with sufficient electrical outlets for 12 microprocessors. The second episode requires two classrooms, one for the computers (electric outlets) and one for apparatus.

The third episode, at MEPS, requires one large classroom with desks and one room with tables and chairs and outlets for the computers. The fourth episode requires the same facilities as the first episode.

Finally, to collect data on the clarity and acceptability of measures, test administration staff will administer brief feedback forms to the subjects after each testing session, as appropriate, as well as conduct post-session interviews with random samples of the subjects (about 5-10 percent). The feedback forms and interviews will focus on the clarity of test items, formats, and instructions; perceived "validity" and "fairness" of the tests; objectionable items, etc. Each team will write a report outlining

its finding from these investigations. We have inquired at FORSCOM about the feasibility of the activities outlined above and have been informed that they could be accomplished.

Subtask 8: Analysis of Trial Battery Pilot Test Data

Rationale. This subtask involves the performance of the analyses of the pilot test data. Therefore, these analyses are designed to answer the three primary questions of concern that we outlined in the Rationale for Subtask 7 above. In the interest of brevity and to avoid redundancy, we refer the reader to that section.

Procedures. Analyses of the predictor responses will include item analyses; test score distributions (frequency distributions, mean, standard deviations, skew, kurtosis, etc.); internal consistency and test-retest reliabilities. Item and test score differences for the major ethnic/sex groups will be examined. Item factors analyses may be performed for predictors that have unknown or ambiguous factors structures. Correlations between predictor measures, including ASVAB scores, will be computed and factor analyses performed as appropriate. As noted earlier, in Subtask 4, the inclusion of "marker" variables in the Trial Battery will make factor analyses more easily interpretable and allow a better understanding of newly developed measures. All these results will be used to identify deficiencies in items and tests such as poor score distributions, low reliabilities, redundancy in the battery, and race or sex differences in item and score distributions. (We should note that we have been informed that

it is not possible to request specific numbers by race and sex, so some of these analyses may not be possible at this stage of the research.)

We will assess the "fakeability" and "practice" effects by comparing score distributions between the experimental and control groups, i.e., the "fake good" versus "fake bad" versus "regular" samples and the "several trials" versus "single trial" groups. T-tests and analysis of variance or analysis of covariance, if appropriate, will be the analytic method. Suspect tests and scales will be identified through these analyses and scrutinized for improvement and deletion. In addition, special scales will have been constructed to detect such response biases, and they will be evaluated to determine whether or not they are indeed performing that function.

Finally, relationships of the Trial Battery measures to available criterion information such as the soldiers' performance in training and performance ratings will be analyzed.

A primary objective of these analyses is the identification of redundancy in the Trial Battery, both within the battery itself and between the Trial Battery and current measures. This information is of particular interest for guiding revisions to the Trial Battery (which takes place in Subtask 10). Assuming computer-assisted measures were administered in Subtask 7 (recall that Subtask 6 includes a decision point about the inclusion of such measures in the pilot test), then we will also prepare a special report about the Trial Battery experience with these measures,

especially the perceptual/psychomotor computerized measures. This report will also address the cost implications of further development efforts, especially costs associated with large-scale deployment and utilization of computer hardware that would be necessary in Subtask 11, administration of the revised trial battery. The availability of sufficient hardware, through Army or contractor resources will, no doubt, play a major role in decisions made about these measures at this point in the project.

As we noted in the overview section, Task 2 staff will have primary responsibility for conducting the analyses outlined here. All the data will be added to the LRDB by Task 1 staff, and they will consult with Task 2 staff on the conduct of the analyses and provide appropriately constructed data files for the analyses.

Subtask 9: Analyze Preliminary Battery: FY83/84 Cohort School and Preliminary Battery Data

Rationale. The purpose of this subtask is to analyze the relationships between the measures on the Preliminary Battery, current pre-induction predictors, and training school performance. The results of these analyses are input to Subtask 10 to guide revisions to the Experimental Battery. Also, we will perform analyses on data collected during the first two or three months of Subtask 3 with the Preliminary Battery in order to provide guidance for development of measures in Subtask 6.

Procedures. Approximately 13,500 soldiers, an average of 3,350 from each of four MOS, will have completed the Preliminary Battery (see Table 3, page 19). Also available will be measures of their performance in training (from Task 3) and their scores on current pre-induction measures (from Task 1). The analyses will be of four major types.

1. Covariances of Preliminary Battery measures -- correlation matrices and factor analyses will be completed to identify redundancies with the battery itself. As noted above, these analyses will be performed on part of the sample to provide guidance for Subtask 6.

2. Covariances of Preliminary Battery measures and current pre-induction measures -- correlation matrices and factor analyses of the two sets of measures will be completed to identify redundancies across the two sets of measures. Again, early analyses on part of the sample will be performed.

3. Prediction of training school performance--bivariate correlations between training measures and Preliminary Battery measures, and between training and current pre-induction measures will be completed. Multiple regressions including current and Preliminary Battery measures will be completed to identify the amount of incremental validity contributed by Preliminary Battery measures.

4. Classification analyses -- the sample is composed of soldiers from four MOS of relatively different occupational types: Radio TT

Operator (05C), Admin Specialist (71L), Vehicle and Generator Mechanic (63B), and Tank Crewman (19E/K). This provides the opportunity, at a minimum, to examine score differences between MOS groups on the Preliminary Battery measures in order to examine their value for classification purposes. More sophisticated analyses, such as multiple discriminate functions analyses or centour analyses are certainly possible given the sample sizes. However, we must keep in mind that these classification or group memberships have not been made optimally, so caution must be exercised. An interesting possibility, however, would be to identify "outliers" within these groups and examine their score profiles. These outliers will be particularly interesting to follow-up in terms of tenure and job performance as those data become available (in Subtask 12).

The overall goal of these analyses is to identify measures or constructs in the Preliminary Battery that are efficient (correlate the least with other pre-induction measures), effective (in this case, predict training performance), and provide incremental validity (beyond that produced by the ASVAB).

Task 2 staff will have primary responsibility for all the analyses outlined above, but Task 1 will have major input to the analyses outlined in points 3 and 4 above. Task 1 staff will add these data to the LRDB and will provide appropriate data files for analyses. This latter point will be especially crucial for the early analyses of the Preliminary Battery.

Subtask 10: Prepare Revised Trial Battery for FY83/84 Cohort Predictor/Performance Data Collection

Rationale. In this subtask, information from the Preliminary Battery analyses and analyses of the pilot test of the first version of the Trial Battery will be integrated. This is essential for guiding revisions to the Trial Battery. Other steps in this subtask are obviously necessary, that is, actually revising the Trial Battery, preparing administration materials, and training test administrators.

Procedures. Analyses from Subtasks 8 and 9 will be used to guide revisions to the Trial Battery so that we have the most efficient, effective and practically feasible set of measures. Task 2 staff will carefully examine the results of these analyses and prepare a revision plan. This plan will be submitted to ARI for review. We anticipate extensive interaction, i.e., meetings in Minneapolis and/or Washington, will be required to make final decisions about the revisions to be made.

The Trial Battery measures will then be revised in accord with these plans. We think the objective should be a Trial Battery that requires a maximum of four hours to administer. This should make it feasible to administer all the trial predictors to each member of the FY83/84 cohort that will be tested on the next subtask. This is desirable in order to have complete data on each subject for analytic purposes and to simplify the data collection procedures, which must include the collection of job performance criteria as well as the Trial Battery measures.

It may be necessary to hold some very small sample try outs of revised measures after revisions have been made to be sure that administration procedures and time estimates are all in order. Our best estimate is that three groups of 25 soldiers for one-half day each would be the maximum requirements. Final administration procedures will then be designed. Testing materials in the quantity necessary will be printed or procured. Administration manuals with detailed instructions will be written.

Data collection teams made up of staff from Task 5 will be responsible for collection of data from the FY83/84 cohort, including the Trial Battery data. Task 2 staff will be responsible for providing the testing materials, detailed administration procedures, and a training session for test administration. We will schedule the training session for the most convenient site, probably Washington, and train all members of Task 5 staff that will perform Trial Battery data collection.

An important part of this training and the content of administration manuals will include procedures to follow, given the occurrence of anticipated problems. The pilot test, conducted by Task 2 staff, should provide sufficient information to prepare such contingency plans for almost all data collection problems.

Subtask 11: Monitor/Assist Administration of Revised Trial Battery to FY83/84 Cohort

Rationale. At this point, prior subtasks have resulted in the development of a Trial Battery of new pre-induction predictors, materials have been

prepared, and the Task 5 staff responsible for administration of the battery have been trained. Task 2 staff, however, will monitor the administration process and be prepared to offer assistance in overcoming administration problems. The Trial Battery will be administered to about 500 soldiers in each of 19 MOS (for which such numbers are available--see Table 3), and job performance criterion data will be collected for these same soldiers. This provides a data set for a concurrent validity study of the Trial Battery. To the extent possible, this sample of soldiers should be stratified on race and sex within MOS. (Some MOS will have no females.) This stratification is necessary in order to carry out studies of test fairness. Also, this sample will include as many soldiers as possible that completed the Preliminary Battery in Subtask 3 during their AIT. This will allow a predictive validity investigation of the Preliminary Battery.

In addition to the collection of these primary data, a number of experimental projects will require the collection of data on some smaller samples. This research will focus primarily on the extraneous factors that might effect the generalizability of the measurement process, that is, practice effects, subject condition effects, faking, etc., and will be very similar to the research carried out in the pilot test. Indeed, the exact nature of these projects depends very much on the outcome of the research on motivation, practice effects, and fakeability conducted in Subtask 8. For present estimation purposes, we assume that four research projects will be required: test-retest reliability (stability) of measures, practice effects for a selected subset of measures, fakeability, and the differences in scores achieved by the primary body of soldiers and scores

achieved by soldiers at an early point in their career, i.e., at AIT. For the stability research project, a sample of 500 will yield stability coefficients with a standard error of about .04, which is sufficiently precise. These soldiers will be a sub-sample of the primary sample, and will complete the predictors of interest 30 days after their first completion. For the research on practice effects and fakeability, separate samples (other than the primary sample) will be required. If we make the same assumptions as outlined in Subtask 8, three samples of 115 soldiers will be required: one for practice effects, one for fake "good," and one for fake "bad." If different assumptions are made, and the outcome of the pilot test may dictate such assumptions, then these estimates will change. One change that is perhaps more likely than others would be an examination of the interaction of sex and/or race with practice or fakeability. If this does occur, then the required sample size will be much larger, i.e., if four groups (black and white females and males) are of interest, then the required sample size would be 320. The fourth investigation of score differences between "early career" soldiers and the primary sample (later career soldiers) must be of sufficient size to provide stable estimates of covariance so that the Trial Battery factor structures can be compared. This is necessary in order to evaluate the extent to which maturation affects the structure of scores on the Trial Battery so that in turn, the limits to be placed on the concurrent validity results can be estimated. A sample size of 1,000 new recruits will provide sufficient stability for these analyses (standard error of correlation coefficient of .03).

Finally, a few comments must be made about the limitations, or anticipated problems due to use of computer-administered measures. As already stated, the entire Trial Battery will be targeted to take no more than four hours, including computer-administered and apparatus measures. The actual mechanics of getting soldiers to complete these measures are not only time bound, they are also constrained by the number of computers and sets of apparatus available at each testing site. As we menitoned in Subtask 7, we estimated that the use of 12 computers would allow the processing of 48 soldiers per day for a two-hour computer battery. If the version of the Trial Battery used at this stage requires one hour, then 96 soldiers could be processed per day--which would possibly be fast enough to keep pace with the administration of the other parts of the prediction battery. But, these data will be collected at any one time at multiple sites, probably as many as 8. Thus, 96 microprocessors could be required to have all subjects take these measures, assuming all that we have just stated. At any rate, information about this matter will be the focus of analyses and a special report at an earlier part of the project (see Subtask 8), so that an informed decision about the inclusion of computerized measures at this stage can be made.

Procedures. Task 2 staff will be "on call" to answer questions from the data collection teams (made up of Task 5 staff) about the main data collection efforts throughout the period of Trial Battery administration. In addition, they will collect the data for the first three of the four special projects outlined above and will train Army personnel at the appropriate TRADOC sites to collect the data for the fourth sample of 1,000 recruits. The nature of the facilities required will be the same as

outlined for the pilot test (classrooms with desks or tables and chairs, cleared classrooms for apparatus, and electrical outlets in rooms where computer tests are given). The number of such rooms at each site will depend on the number of soldiers to be tested and the period of time allowed to collect the data.

Completed test materials will be sent directly to Task 1 staff for addition to the data base.

Subtask 12: Analyze FY83/84 Cohort Data: Trial Battery/Performance Measures

Rationale. This subtask involves the analysis of the data collected in the previous subtask, in order to guide the preparation of the Experimental Battery (Subtask 13) for administration to the FY86/87 cohort (Subtask 14). There are two major analytic efforts: a concurrent validity investigation of the Trial Battery and a predictive validity investigation of the Preliminary Battery.

Throughout these subtask procedures, Task 1 and 2 staff will work closely. Task 2 staff will bear primary responsibility for the analyses mentioned in points 1 and 4 below, while Task 1 staff will bear primary responsibility for points 2, 3, and 5 below and will provide appropriate data files for all analyses.

Procedures. We first discuss the concurrent validity analyses. These fall into five general categories.

1. Reliability/Generalizability. We are concerned here with the internal structure of each measure and the extent to which the observed score on a measure is affected by extraneous factors, or the extent of generalizability of observed scores.

   The internal structure of each measure can be examined by the usual internal consistency estimates (coefficient alpha or KR20). Where these estimates are lower than desired (say, less than .85), there may be multidimensionality in the item set. In such cases, factor analyses of the items may be used to identify unidimensional subsets of the items. Assistance from the Task 1 team will be used at this point so that the most appropriate factor analysis methods are used. We should point out here that unidimensionality is desirable, but will not be pursued to the detriment of achieving validity in predicting job performance. As members of the Scientific Advisory Group have pointed out, multidimensionality is not necessarily a problem for a predictor.

   Latent trait methods may then be used to calibrate items within the unidimensional item sets, although more traditional methods of item analysis will also be used, especially for predictors where unidimensionality is not as important.

The effects of extraneous factors on observed test scores will be investigated by using the generalizability theory approach. Data will have been obtained on order of administration, practice, time (i.e., test-retest), and fakeability for those variables for which it is appropriate. For example, fakeability data will be collected for the bio/vocational interest inventory, but not for paper-and-pencil cognitive tests, and practice data will be collected for psychomotor effects, but not for the bio/vocational interest inventory, etc. If possible, balanced factorial designs will have been used to collect these data wherever interactions between effects are hypothesized. For most variables, however, we think no interactions can reasonably be assumed. This will allow us to compare the scores for persons under a given experimental condition to a large "normal" group; that is, to carry out independent, single effects analyses of practice, time, etc.

Where extraneous factors are found to affect observed scores, test procedures or content will need to be evaluated with regard to changes that may obviate such effects. For example, where practice effects occur, testing procedures will need to be arranged to provide a longer "warm-up" or unscored testing time. If faking appears to affect scores, items must be appropriately weighted to detect the type and direction of faking, and instructions for tests or inventories altered to prevent faking.

2. Validity/Fairness. As we said, the FY83/84 cohort data will provide the opportunity for concurrent validity analyses. To review, criterion data will be available from measures developed in Tasks 4 and 5. We will have the ASVAB scores and other pre-induction measures for the FY83/84 cohort in the longitudinal data base. Finally, we will have Trial Battery data from 1,000 new recruits collected coincident with the collection of data on the FY83/84 cohort.

All these data will be used to estimate the validity and fairness of the new predictor measures. The overriding objective of these analyses is to identify the most efficient set of new predictors that increases the accuracy of prediction of soldiers' job performance in a manner that is fair for race/sex subgroups. Classification of soldiers is also an important objective, but is dependent upon the identification of predictors that add to accuracy of prediction. (It is also the case, however, that classification considerations will have an important effect on decisions about which new predictors should be retained for further investigation, since a predictor may add little to a general prediction equation, yet still be very useful in differentiating success in different occupations.)

There are at least two basic approaches to reducing the size of a predictor battery. The first is an internal structure approach, and the second an external validity approach. In the first approach, we

choose the subset of predictors that: (a) have high internal consistency reliabilities for each measure, and (b) have very low correlations between measures. In the second approach, we choose those predictors that: (a) have high correlations with external criteria of interest, and (b) minimal correlations with each other. Both of these approaches will be applied at the item level or test (scale) level. Furthermore, tests or scales will be rescaled so that nonlinear test scoring methods can be evaluated. Finally, items or tests that appear useful with regard to the above internal structure and/or external validity criteria must be evaluated with regard to fairness for various subgroups. The ideal predictor item, then, will have the following characteristics: (a) high correlation with the predictor scale it purports to measure; (b) low correlation with other predictor scale scores (and items in those scales); (c) high correlation with Army success and/or job performance criteria; and (d) similar response characteristics and relationships with external criteria across race and sex subgroups; that is, "fair." The ideal predictor scale will have similar characteristics: (a) high internal consistency, (b) low correlations with other predictor scales, (c) high correlations with Army success and/or job performance criteria, and (d) fairness.

There are many statistical methods available for use in achieving the above analysis objectives, and we will work closely with Task 1 staff to identify and use those methods most appropriate for the particular analysis problem. (Some of these techniques are explanatory and confirmatory factor analyses, multiple regression and other prediction

optimization algorithms; Cleary, Thorndike, etc., models of test fairness.) We should also note that we will not duplicate analysis efforts performed by Task 1 staff; indeed, we envision very close cooperation in the planning of analyses.

As noted above, a major limitation of the inferences to be made from analysis of the FY83/84 cohort arises from the fact that it is a concurrent validity design. About 30 percent of the cohort may have attrited, and those remaining will have had many months of Army experience, including training designed to improve the soldiers' skills in areas appropriate to overall soldier performance and specific job performance. As noted earlier (p.2-76), we will have trial battery predictor data available for 1,000 new recruits. These data will enable us to estimate the effects of restriction of range and changes in the factor structure of the Trial Predictor Battery due to attrition and experience.

3. Utility. Information will be provided about the utilities of various levels of performance in various MOS so that we may more adequately evaluate the incremental validity and utility of the predictors. (Scaled utility values will be obtained by Task 4 staff.) Therefore, a straightworward analysis of incremental validity, such as increments in R2 when new predictors are added to currently available predictors, must be informed by, and coordinated with, the utility analyses of Task 1. There are basically two questions to be answered: (a) Does a

new predictor(s) increase predictive accuracy over that available with current predictors; and (b) if so, by how much does the utility increase exceed any additional costs of recruitment, assessment, and induction?

4. Investigation of Predictor Scales. A related set of analyses concerns the investigation into the psychometric and psychological meaning of various intervals on the new predictor scales. Each new predictor scale will be investigated to find out (a) how persons falling in each quintile, for example, score on other predictors and on various criterion measures; and (b) ceiling effects, floor effects, and changes in the error of measurement across quintiles. Of course, examinees can be grouped into either finer or coarser gradations as indicated by the data.

5. Predictive Validity Analyses. These analyses will be used as a point of comparison to the concurrent validity analyses. It is extremely likely that many constructs, if not actual measures, will be in common between the Trial Battery used in the concurrent validity investigation and the Preliminary Battery. In fact, as we earlier remarked in Subtask 4, we will attempt to have some of the same scales in both batteries. Thus, we will have available concurrent and predictive validity coefficients for some subset of the constructs measured in the Trial Battery. In addition, a subset of the sample will have completed the Preliminary Battery when they were in AIT (Subtask 3)

and the Trial Battery (in Subtask 11) and will have job performance criterion scores on record. Although maturation effects make interpretations nonstraightforward, we will thus have empirical correlations between scores on the Preliminary and Trial Battery measures. The actual predictive validity analyses, of course, will consist primarily of correlations between the Preliminary Battery measures (administered in Subtask 3) with the criterion data collected in Subtask 11.

These two sets of analyses, e.g., the predictive and concurrent validity analyses, will be integrated and a report of findings will be submitted to ARI, with suggestions for revisions of the Trial Battery. It is our belief that these suggestions will consist primarily of reductions in the battery, at either or both the item or scale level. This report will be discussed with ARI in order to make final revision plans. Draft and final field test plans for the FY86/87 cohort data collection plan will be prepared after the revisions have been approved by ARI.

Subtask 13: Prepare Experimental Battery for Administration to FY86/87 Cohort

Rationale. The purpose of this subtask is to make the revisions to the Trial Battery that were decided upon in the previous subtask, i.e., prepare

the Experimental Battery, prepare associated administration materials, and train the personnel that will be administering the battery.

This final version of the Experimental Battery will be administered to an average of 2,200 recruits as they enter AIT for each of the MOS selected as the focus of the project. As noted in the Introduction to this plan, the rationale for this sample size is to provide a sufficiently large longitudinal sample for predictive validity analyses of the Experimental Battery, given anticipated rates of sample attrition. The experience gained in all prior battery administrations will be used to prepare for this administration, especially the administration of the Preliminary Battery to the FY83/84 longitudinal sample, since it will also have taken place at AIT sites which will be involved in the FY86/87 cohort administration.

Procedures. After final revisions are made, sufficient quantities of the printed portions of the battery will be procured. Then, apparatus and microprocessors must be obtained as previously noted. Detailed administration manuals will be written and a one-to-two day training program developed. Site visits of approximately one week duration will then be made to the TRADOC posts where data collection will occur. Persons selected for data administrators will be trained, facilities inspected, and the apparatus and computers put in place.

The exact details of the steps outlined in the above paragraph depend entirely on the final contents of the Experimental Battery. At this point, we think it will consist of a two-to-three hour battery that will be a mixture of paper-and-pencil and computer-administered tests. As much as possible, computer-administered testing will be used in order to reduce testing time. This implies that the administrators must be experienced enough or receive sufficient training to handle computer administered tests. Local, on-site administrators will be required for the same reasons as were cited for the Preliminary Battery administration (i.e., infeasibility of Task 2 staff being on site over the entire one-year period that data are collected--see Subtask 3). By this point in the project the administration procedures should be very well honed and we should be able to train test administrators for virtually all problems or contingencies. Also, just as for the Preliminary Battery, one person will be designated as the primary contact and will be responsible for security of testing materials, shipping completed batteries, monitoring attendance at testing sessions, and communicating with Task 2 staff.

Finally, we note that we are currently conducting site visits at TRADOC posts that would likely be involved in the test administration and are informing them of the nature of the demands on soldier time and the need for administration personnel. To date, the information obtained from these visits indicates that the plans can be carried out.

Subtask 14: Monitor Administration of Experimental Battery to FY86/87 Cohort, Further Analyses of FY83/84 Cohort Data

Rationale. Although Task 2 staff will have trained local, on-site personnel and they will be administering the Experimental Battery, close monitoring of the administration process will be required. Given the somewhat limited time that will have been available for analyses of the FY83/84 cohort in Subtask 12, further analyses will also be completed during this subtask.

Procedures. Task 2 staff will make several scheduled and unscheduled visits to each test site to observe test administration, test security procedures, and to address any problems that occur during the administration process, i.e., one entire year, from March, 1986 through February, 1987. We will also set up a regular, by-phone reporting procedure after every weekly administration. (The battery will be administered, if possible during the first week of the soldiers' AIT in order to reduce the effects of training on Experimental Battery scores.)

We will monitor the numbers of soldiers tested so that progress toward target sample sizes can be tracked. It is probable that testing will not be required for every class in all 19 MOS and this tracking process will be the means by which the actual administration sessions are controlled. Finally, with regard to monitoring, Task 2 staff will be continuously available to answer questions (via phone) or to make short notice visits in response to problems.

The exact nature of the further analyses of FY83/84 cohort data are dependent upon what has been accomplished earlier. We note here that any loose ends will be tied up and interesting further analyses will be pursued at this time.

Subtask 15. Analyze FY86/87 Cohort Data and Prepare Final Reports

Rationale. Task 2 concludes with this subtask. The purposes of the subtask are to compare the FY83/84 cohort and FY86/87 cohort in terms of their success on Experimental Battery measures, analyze the covariance of the final predictor battery measures within itself and with then current pre-induction measures, and analyze the relationship of the final Experimental Battery to training performance measures. The ultimate goal of these analyses is to identify and recommend the best battery for operational use, based on all data at hand. (These recommendations will be subject to later revision, however, since a follow-up predictive validity investigation of the final Experimental Battery will be completed by when second tour performance measures are available for the FY86/87 cohort.)

Procedures. The following sets of analyses will be carried out:

1. Comparison of FY83/84 and FY86/87 cohort data. The primary foci of this analysis will be range restriction, factor structure, and psychometric/psychological meaning of the Experimental Battery. Recall that the final battery administered to the FY86/87 cohort will be a subset of that administered to the FY83/84 cohort. Therefore, the batteries will be different in the contextural sense; that is, the

FY83/84 cohort will have completed more predictors, slightly longer predictors, and perhaps, in a different order. This somewhat limits the interpretations that can be placed on comparisons of the two sets of data, but not unduly so. It will still be the case that the FY86/87 cohort battery will be a subset of the FY83/84 battery, so equivalent sets of tests and/or items for both samples can be assembled.

Item and scale distributions will be computed and compared to identify range restriction effects. (Task 1 researchers could then use these data to refine the earlier FY83/84 cohort analyses; i.e., assemble scores on the reduced predictor battery and correct the relationships of these scores with job performance criteria for restriction of range.)

Confirmatory factor analysis techniques will be used to see if the factor structure of the new predictors on the FY83/84 cohort applies to the FY86/87 cohort. If not, then factor analyses will be done to identify the new factor structure differences. (A major concern will be the attribution of factor structure differences. Are they due to true cohort difference, FY83/84 vs. FY86/87 recruits, or due to attrition and experience--which is present in the FY83/84 cohort, but not the FY86/87 cohort? Recall that data, collected concurrently with the FY83/84 cohort data, will be available from a sample of 1,000 new recruits. These data will be of obvious usefulness for probing this question.)

Finally the FY86/87 cohort data will be analyzed to discover the psychological/psychometric meaningfulness of the predictors and various scale intervals on the predictors. These analyses will benefit from prior similar analyses of the FY83/84 cohort data, which will provide direction for these analyses. The investigation of factor structure will also inform these analyses and aid in focusing this effort.

2. Relationship to training criteria. Training criteria data will be available for the FY86/87 cohort (from Task 3). The relationships of the experimental battery to these criteria will be thoroughly investigated. These analyses will focus on the absolute and incremental validity of the Experimental Battery for training completion and success, although another interesting problem is the prediction of success at various stages of training. If appropriate training criteria are available, these kinds of analyses will be completed.

3. Covariances of Experimental Battery measures. Correlation matrices and factor analyses will be completed to identify redundancies within the battery itself.

4. Covariances of Experimental Battery measures and current pre-induction measures. Correlation matrices and factor analyses of the two sets of measures will be completed to identify redundancies across the two sets of measures.

5. <u>Classification analyses</u>. The FY86/87 cohort sample will be composed of soldiers from MOS of different occupational types, intended to represent the diversity of all Army jobs. This provides the opportunity, at a minimum, to examine score differences between MOS groups on the Experimental Battery measures in order to examine their value for classification purposes. More sophisticated analyses, such as multiple discriminate functions analyses or centour analyses are certainly possible, given the sample sizes. However, we must keep in mind that these classifications or group memberships have not been made optimally, so caution must be exercised.

An interesting possibility, however, would be to identify "outliers" within these groups and examine their score profiles. These outliers will be particularly interesting to follow up in terms of tenure and job performance as those data become available. Draft, and after-ARI review, final, technical reports will be prepared on the final set of recommended instruments and on all technical work performed on Task 2.

SUMMARY OF EXPECTED OUTCOMES FROM TASK 2

Operational Outcomes

1. Non-cognitive Attributes Inventory. This will be a relatively short, untimed, paper-and-pencil inventory suitable for administration at Military Entrance Processing Stations. The inventory will contain the most efficient set of measures of biographical data, and vocational interests that proves useful for the selection and classification of applicants. Scores on inventory scale will be input to the selection and classification algorithms. Another possible use of this inventory is its administration at recruiting stations. Recruiters could use scores generated from the inventory to counsel recruits in their choice of MOS. This latter use is especially feasible if the capability for computerized administration and scoring is in place, which would go a long way toward overcoming "test security" problems. Although the bulk of the research with the inventory will most likely be conducted in a paper-and-pencil format, this instrument would be very amenable to conversion to a computer-administered format, and research will have been conducted to determine the comparability of results across these two formats.

2. Perceptual/Psychomotor Battery. This will be a battery of measures in the perceptual/psychomotor area that will be primarily, if not completely, computer administered. The measures will tap constructs that have shown to tap unique variance over and above that measured on the ASVAB. (At this point in time, the major unresolved issue is whether large-scale data can be collected on these computer-administered measures, given practical constraints of time and money. We have assumed that positive decisions were reached at the various decision points outlined in this research plan.)

3. Another operational outcome will be information about the vulnerability of the set of non-cognitive measures to differing motivational sets (comparisons of responses of soldiers on active duty to applicants at MEPS) and to faking. This information should enable the Army to make informed decisions about the reliance that can be placed on these measures in an operational setting.

4. Additional Cognitive Measures or Improved Cognitive Measures. These will be paper-and-pencil measures of cognitive abilities not presently measured or improved versions of those currently in use.

5. In a sense, all the outcomes listed above can be thought of as optimistic. These outcomes assume that the research will show that

new pre-induction measures will show incremental validity (over the current measures) and/or increase the efficiency of classification of recruits into MOS. The research may show, however, that some or all of the new measures do not result in such increments. We maintain that this information is a highly valuable operational outcome, since it would confirm that current pre-induction measures would be effectively performing the job of selection and classification.

Scientific Outcomes

1. Delineation of empirical relationships between measures of human attributes across major domains. Although relationships between measures within major domains (e.g., within traditional cognitive tests) have been fairly well mapped out, there is much less information available about relationships between measures from different domains. Task 2 research will provide such empirical information.

2. Tests of Validity in Applied Settings. Several of the measures that will likely be used in this project have relatively little available evidence of their validity in the applied setting; i.e., for predicting succes in training and on the job. Interest measures have been shown to predict occupational entry and longevity but have been less well researched with respect to degree of successful job performance.

The newer cognitive/perceptual measures have generally not been evaluated in applied settings (Hunt, 1983), and to a lesser extent, this is true of and psychomotor measures. Task 2 research will provide a rigorous investigation of the "applied validity" of such measures.

3. Incremental Validity. The points made just above apply equally well to the question of incremental validity. Task 2 should be able to provide a definitive answer to the question of how much increase in the accuracy of prediction for the job performance in disparate jobs can be obtained by adding some non-cognitive measures, perceptual/psychomotor measures, or additional cognitive measures to the ASVAB--which is an excellent representative of traditional cognitive tests used to predict training and job performance.

4. Linear Composite vs. Subgroup Approaches to Selection/Classification. Owens and Schoenfeldt (1979) have championed a subgrouping approach to the problem of prediction in contrast to the more commonly employed approach of linear composites. Very briefly, the subgrouping approach advocates the classification of persons into one of a finite set of groups, based on scores on a set of measures, and then making similar predictions for those individuals in the same group. The linear composite approach advocates measuring persons on several measures, and then applying a set of linear weights to a person's scores on those measures to make predictions. Task 2 will provide sufficient measures to operationalize both methods and compare their effectivenes in selection and classification.

5. Validity Generalization. Recent research has shown that the validities of cognitive tests generalize quite well across different kinds of settings and occupations (Schmidt, Hunter & Pearlman, 1981; Schmidt & Hunter, 1978; Schmidt, Hunter, Pearlman & Shane, 1979). Little or no research exists, however, on the degree to which other types of predictors generalize. Task 2 provides the opportunity to extend the investigation of validity generalization to these other types of predictors. Measures of biographical data, vocational interests, perceptual, and psychomotor abilities will be administered to soldiers in a variety of MOS, and school, attrition and job performance criteria data will be available. Validity generalization analyses will be conducted for all predictor measures, as well as predictor composites. These findings should significantly contribute to the growing body of knowledge about validity generalization.

REFERENCES

Bowmas, D.A., & Heckman, R.W. Job analysis of the entry-level firefighter position. Minneapolis: Personnel Decisions, Inc., 1976.

Brogden, H.E. Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. Educational and Psychological Measurement, 1951, 11, 173-196.

Callender, J.C., & Osburn, H.G. Testing the constancy of validity with computer-generated sampling distributions of the miltiplicative model variance estimate: Results for petroleum industry validation research. Journal of Applied Psychology, 1981, 66, 274-281.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Dunnette, M.D. Basic attributes of individuals in relation to behavior in organizations. In M.D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally, 1976.

Dunnette, M.D., & Borman, W.C. Personnel selection and classification systems. In M.R. Rosenzweig & L.W. Porter (Eds.), Annual Review of Psychology. Palo Alto, CA: Annual Reviews, Inc., 1979, 30, 477-525.

Dunnette, M.D., Rosse, R.L., Houston, J.S., Hough, L.M., Toquam, J., Lammlein, S., King, K.W., Bosshardt, M.J., & Keyes, M.A. Development and validation of an industry-wide electric power plant operator selection system. Minneapolis: Personnel Decisions Research Institute, 1981.

Guion, R.M. On trinitarian doctrines of validity. Professional Psychology, 1980, 11, 385-390.

Hunt, E. On the nature of intelligence. Science, 1983, 219,141-147.

Hunter, J.E., Schmidt, F.L., & Jackson, G.B. Integrating research findings across studies. Unpublished paper in Methodological innovation in studying organizations. Symposium presented at the Center for Creative Leadership, Greensboro, NC: 1981.

Linn, R.L., Harnish, D.L., & Dunbar, S.B. Corrections for range restriction: An empirical investigation of conditions resulting in conservative corrections. Journal of Applied Psychology, 1981, 5, in press.

Maier, M.H. Validation of selection and classification tests in the Army (Working Paper: Personnel Utilization Area 82-2). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, 1981.

Messick, S. Test validity and the ethics of assessment. American Psychologist, 1980, 35, 1012-1027.

Owens, W.A., Jr. Background data (Chapter 14, pp. 609-645). In M.D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago, IL: Rand McNally, 1976.

Owens, W.A., & Schenfeldt, L.F. Toward a classification of persons. Journal of Applied Psychology, 1979, 64, 569-607.

Pearlman, K., Schmidt, F.L., & Hunter, J.E. Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. Journal of Applied Psychology, 1980, 65, 373-406.

Peterson, N.G., & Bownas, D.A. Task structure and performance acquisition (Chapter 3). In M.D. Dunnette & E.A. Fleishman (Eds.), Human capability assessment. New York: Lawrence Erlbaum & Associates, 1982.

Peterson, N.G., & Houston, J.S. The prediction of correctional officer job performance: Construct validation in an employment setting. Minneapolis: Personnel Decisions Research Institute, 1980.

Peterson, N.G., Houston, J.S., Bosshardt, M.D., & Dunnette, M.D. A study of the correctional officer job at Marion Correctional Institution, Ohio: Development of selection procedures, training recommendations and an exit information program. Minneapolis: Personnel Decisions Research Institute, 1977.

Peterson, N.G., Houston, J.S., & Rosse, R.L. The LOMA job effectiveness prediction system, technical report #4: Validity analyses. Atlanta, GA: Life Office Management Association, in press.

Schmidt, F.L., & Hunter, J.E. Moderator research and the law of small numbers. Personnel Psychology, 1978, 31, 215-231.

Schmidt, F.L., Hunter, J.D., & Caplan, J.R. Validity generalization results for two job groups in the petroleum industry. Journal of Applied Psychology, 1981, 66, 261-273.

Schmidt, F.L., Hunter, J.E., & Pearlman, K. Task differences as moderators of aptitude test validity in selection: A red herring. Journal of Applied Psychology, 1981, 66, 166-185.

Schmidt, F.L., Hunter, J.E., Pearlman, K., & Shane, G.S. Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. Personnel Psychology, 1979, 32, 257-281.

Walker, H.W., & Lev, J. Statistical inference. New York: Henry Holt & Co., 1953.

Wernimont, P.F., & Campbell, J.P. Signs, samples and criteria. Journal of Applied Psychology, 1968, 52, 372-376.

TASK 3 RESEARCH PLAN

MEASUREMENT OF SCHOOL/TRAINING PERFORMANCE

GENERAL PURPOSE OF TASK 3

The general purpose of Task 3 is to generate information about the performance of soldiers in training, to be used in the validation of initial predictors and in the prediction of first-tour and second-tour performance in the Army. To accomplish this purpose, existing measures of training performance will be analyzed and evaluated, new measures will be developed where neeeded, and composite sets of predictor and criterion measures will be assembled.

As job performance surrogates, training measures can serve to reduce the time required for predictor validations from years to months. When used to predict subsequent performance, training measures have the potential to increase the accuracy of classification into MOS over that obtained by the use of pre-induction predictors alone. Both the extent to which training measures can be used as surrogates for more ultimate job performance criteria and the degree of incremental validity obtained by including training success as a predictor itself will be assessed in Task 1.

A further purpose of Task 3 is to collect and interpret training performance data derived from recent and current measures and to enter these data into the Longitudinal Research Data Base (LRDB) for use by other tasks. Training performance data from the FY81/82 cohort, for example, will be used by Task 1 to make initial assessments of the ability of

current pre-induction predictors to predict training performance. This procedure is an evaluation of selection tests rather than of training effectiveness.

BACKGROUND ISSUES AND RATIONALE

A principal issue that will be addressed in Task 3 is that of the definition of training success. As explained below, this issue is particularly important because the characteristics to be sought in training measures may differ according to whether the measures are used as administrative criteria, as criteria in predictor validation, or as predictors themselves. This issue also involves a related question: What mechanism explains the predictive relationship between training performance and job performance?

It is naturally desirable to use as reliable and comprehensive measures as possible to obtain training performance information in Task 3. This raises two further issues: (a) how much reliable variance is there in existing (and newly developed) training measures, and (b) what components of training performance, if any, are not currently represented by existing measures? Finally, whether training performance is to serve as criterion or predictor, the additional question remains of how to sample the training performance domain.

Definition of Training Success

The way in which trainee achievement, or success, is conceptualized, defined, and measured is a function of several factors. The major considerations are as follows:

1. What are the overall organizational goals that the training program is intended to serve? For example, is it to produce graduates who can quickly step into a specific job and perform satisfactorily as long as conditions don't change too drastically, or is it to prepare individuals for a very dynamic job environment in which equipment and specific job duties will change considerably over the individual's tour of duty?

2. What model or framework was used to design the training program? (For example, were very specific behavioral objectives used to specify the content? Was the intent to teach fact or skills?)

3. What sources were used to generate the training content? (For example, supervisory complaints, systematic needs analysis of job incumbents, human factor specifications for new situations or equipment, the trainer's theory about what should be taught.)

4. What are the objectives for which the criterion measure will be used? For example:

   - to identify which skill and knowledge areas have been mastered and which need remedial work.
   - to evaluate the strengths and weaknesses of the training program itself.
   - to certify the individual as ready for promotion to the next course or for entry into the job.

Choices among these factors will also have an influence on the degree to which success in training is related to (correlated with) job performance. To the extent that: (a) a training program is meant to serve as certification for entry into the job; (b) the content is derived from a job or task analysis of job encumbents; (c) the training objectives were designed to cover all major job task factors; (d) teaching more general analytic, problem solving, or technical skills is part of the training objectives; and (e) the training program does not control for individual differences in ability, the correlation between training achievement and job success should be maximized.

It follows that if the presence or absence of a correlation between training and performance is to be explained or influenced, the above factors are what must be accounted for. Of particular importance is whether the content of the training criterion is limited to the specific training objectives or whether it is sampled from job content.

This entire issue would be moot if training requirements and job requirements were identical, and to make them as similar as possible is the goal of much of the Army's current training development procedures. But this is a difficult goal, and differences between the behaviors conducive to training success and job success will inevitably exist. Some differences are even inherent in the fact that, to achieve economies of scale, training must be more formal and structured than the job. Those who learn best in one situation may not be those who do so in the other. (Certainly there is anecdotal evidence that many effective job performers were not distinguished academically.)

In order to assemble appropriate predictor and criterion composites of training measures, then, it will be necessary to determine the relation of existing measures to training content and to job content. Likewise, it will be necessary to investigate the mechanism by which training performance predicts job performance, by relating both training-specific and job-specific test items to MOS-specific criteria. To determine the relation of existing measures to training content and to job content, measures will be evaluated at Army schools (Subtask 3.2). In addition, the procedures currently followed to develop Army training content and training measures will be identified (Subtask 3.3). As required, additional job knowledge tests will be developed (Subtask 3.4), and job knowledge items will be identified as school-learned, job-learned, or both (Subtask 3.6).

Reliable Variance

Training is designed to eliminate individual differences by bringing each soldier to the training standard. For improving selection and classification, however, measures are needed that possess substantial reliable variance. For this reason, it will be necessary to review existing measures and examine training courses to seek out components of training performance that exhibit the greatest amount of true variability.

As an example, it is a common practice to allow a trainee several attempts to pass a performance test, without penalizing the soldier for early failures. For administrative purposes, the primary concern is that each soldier reach the standard, not whether one soldier takes longer than

another. For the purpose of developing reliable criterion and predictor measures, however, it would be desirable to refine the scoring procedure for such a measure in order to extract such information as number of attempts to reach mastery and time to reach mastery. In particular, it will be necessary to focus on performance early in training, when the true variability across students can be expected to be greater.

Coverage of the Training Performance Domain

In the past, Army training measurement focused on paper-and-pencil knowledge tests and thus primarily on the cognitive components of the training performance domain. In more recent years, the Army has emphasized hands-on testing, thus capturing perceptual/motor components of the domain as well. This shift toward "performance testing," however, was due principally to the administrative decision that the hands-on format was preferable overall, not to a finding that training success is due almost entirely to motor skill.

It cannot be assumed, then, that the allocation of paper-and-pencil and hands-on tests in a given MOS proportionally reflects particular components of the performance domain. Likewise, there may be components of training success that are not represented by any existing measures. Accordingly, Task 3 will develop: new methods of assessing requirements that are presently difficult to assess, new performance indices, and measures of general performance, as required to represent the domain of training performance as completely as possible. These measures will be developed in Subtask 3.5: Construction of Prototype Measures.

Sampling

Regardless of whether training measures are derived from training or job content, evaluation of training performance entails selecting a basis for sampling the content. The measurement literature provides little systematic guidance for resolving questions about appropriate strategies. The Interservice Procedures for Instructional Systems Development (1975) lists these sampling criteria, but leaves the choice to the user: percent of persons performing, percent of time spent in performing, probable consequences of inadequate performance, task learning difficulty, probability of deficient performance, length of time spent in performing, and length of time between job entry and task performance. The Army in its Guidelines for Development of Skill Qualification Tests (1977) lists the following: known performance deficiencies, tasks contributing to the operation or maintenance of critical combat systems, tasks related to deficiencies in crew or unit performance, tasks that have been revealed as important in prior evaluations, and proportional samples from different content or functional areas of performance. Again the choice of a factor or combination of factors to be used in sampling is left to the user.

One difficulty with factors such as many of these is that they are defined in terms of variables external to job behavior itself. As a consequence, the same behavior evaluated in two different contexts can legitimately be placed in different categories. Although sampling on the basis of such extrinsic factors cannot be avoided entirely, it is clearly desirable to attempt to represent whatever kinds of behavior are present in job performance. Several bases for identifying different types of behavior are

available. The <u>Instructional Quality Inventory</u> (Ellis, Wulfeck & Frederick, 1979) classifies training objectives, test items, and components of instruction in the following categories: fact, concept, procedure, rule, and principle. Lumsdaine (1960) suggested the following categories for the classification of training content: learning identification, perceptual discriminations, comprehension of principles and relationships, procedural sequencing, decision-making, and perceptual-motor skills.

When these and similar classification schemes are examined, three major categories emerge that we consider to have possible implications regarding test format: content that does not require generalization, e.g., the application of a procedure; content that requires generalization, e.g., the application of rules and principles; and content that requires perceptual and/or motor skill. Components of tasks that require perceptual and/or motor skill will become candidates for hands-on testing. Other components of tasks will be classified as requiring/not requiring generalization so that both categories of behavior will be represented in paper-and-pencil knowledge tests. Special attention will be given to performance requiring generalization to determine whether the paper-and-pencil format introduces artificial cues that diminish the applicability of that format for that category of behavior. If so, prototype measures (Subtask 3.5) may be needed to represent content involving generalizing.

A perhaps more difficult question will remain: how to select from within these major categories? We hope to extend the procedures developed by Wheaton, Fingerman, and Boycan (1978) in a rather restricted situation (qualification testing for tank gunnery). In one of the few studies that

has analyzed sampling strategies in relation to testing purpose, Wheaton et al. discussed the suitability of six possible bases for sampling test content from a domain of training objectives: random sampling, frequency of task performance, task difficulty and performance variability, generalizability of objectives, criticality of objectives, and task commonality. They decided to maximize coverage of the job domain by sampling on the basis of task commonality. They selected tasks that had a maximal number of elements in common with other tasks in the domain but a minimal number of elements in common with other tasks on the test. In contrast to factors such as frequency, criticality, and importance, task commonality is not defined in terms of variables external to a task, but simply in terms of number of identical overt behaviors. Although judgments may ultimately be required in establishing commonality, this approach offers possibilities for superior reliability and objectivity.

Sampling on the basis of commonality of task elements will also be employed in Task 3. Since the tasks to be encountered in this work are expected to be more heterogeneous than those dealt with in the Wheaton et al. study, establishing a basis for estimating commonality will require considerable effort. Nevertheless commonality is seen as the most promising approach to the sampling issue.

OBJECTIVES

1. Collect, analyze, and interpret information about existing training criterion measures to augment the FY81/82 cohort data, and provide this information to other tasks. This includes an evaluation of specific measures represented in the Enlisted Master File and recommendations as to what information now collected at schools but not entered in school records and/or entered in the FY81/82 cohort data base should be entered into the LRDB.

2. From the available measures in each of the 19 MOS, assemble the most appropriate set of training performance criteria to be used in validating the selection and classification measures developed by Task 2 and in determining the incremental validity obtained by using training performance in addition to pre-induction information in predicting MOS-specific and Army-wide performance.

3. Determine the extent to which the predictive relationship between training performance and job performance is attributable to content learned in training versus content learned on the job versus general cognitive ability.

4. Advise Army trainers on how existing performance measures and scoring procedures can be refined to increase reliability and amount of information obtainable from training measurement.

5. Construct an end-of-course comprehensive job knowledge test for each of the 19 MOS. These will provide a set of common measures across all MOS in Project A.

6. Develop prototype measures of components of training performance not represented by existing measures or newly developed job knowledge tests.

OVERALL SUMMARY OF THE PROCEDURE

Criterion and predictive information about the performance of persons in training in Army schools will be provided through analysis of performance on recent and currently administered school measures and analysis of performance on improved and newly developed measures. School performance information will be generated in the following subtasks:

Literature review and planning. We have reviewed the literature on the issues and methods of evaluating student achievement and are currently drafting abstracts. We have prepared a master plan for Task 3 and will subsequently submit field test plans as specified in the master plan.

Evaluation of existing measures. The performance of persons in training in the FY81/82 and FY83/84 cohorts will be examined based on studies of information available in school records and already entered in the LRDB by ARI. The tests currently used in Army schools will be examined in discussions with SME to determine the relation of their content to training requirements and will be examined statistically to determine the adequacy of their measurement characteristics.

Analysis of Army training and evaluation procedures. The primary purpose of this subtask is to aid in determining the content validity of current training exercises and training measures by identifying the processes by which these components of the training system are derived and by establishing their relation to job content.

Revision/construction of new comprehensive knowledge tests. To provide improved measures to serve both as criteria of school performance and as predictors of job performance, new comprehensive knowledge tests will be developed in each of 19 MOS. Knowledge will be sampled based on commonality across the MOS tasks, on estimates of frequency of error in performance, and on representation of two classes of task components: those requiring the application of procedures and those requiring the generalization of information.

Development of prototype measures. To represent components of training performance not represented by existing measures or newly developed job knowledge tests, new indices will be derived from existing measures, and other prototype measures will be developed.

Identification of training-relevant and job-relevant test content. To provide a basis for interpreting predictive relationships between the new comprehensive knowledge tests and subsequent job performance, the relevance of the knowledge test items to training and job content will be determined in two ways. First, training relevance will be determined empirically by comparing the performance of entering trainees and graduating trainees; job relevance will be determined by comparing the performance of graduating trainees and job incumbents. Second, job relevance and training relevance will be determined judgmentally by trainers at Army schools.

Develop predictor and criterion composites of school measures. School measures determined to have adequate reliable variance and content validity will be assembled into integrated sets to serve as criteria for validating

initial predictors and as predictors of MOS-specific and Army-wide performance.

Analyze predictive relationships and prepare reports. The results of analyses of FY83/84 and FY86/87 cohorts will be presented in a technical report and an instruments and measures report shortly after the new predictors and new training measures are administered to the FY86/87 cohort.

Table 1 summarizes the subject matter expert (SME) and test subject support requirements for the above Subtasks. Figure 3-1, immediately following, depicts the Task 3 Schedule for the accomplishment of these Subtasks.

TABLE 3-1
SME AND TEST SUBJECT SUPPORT REQUIREMENTS

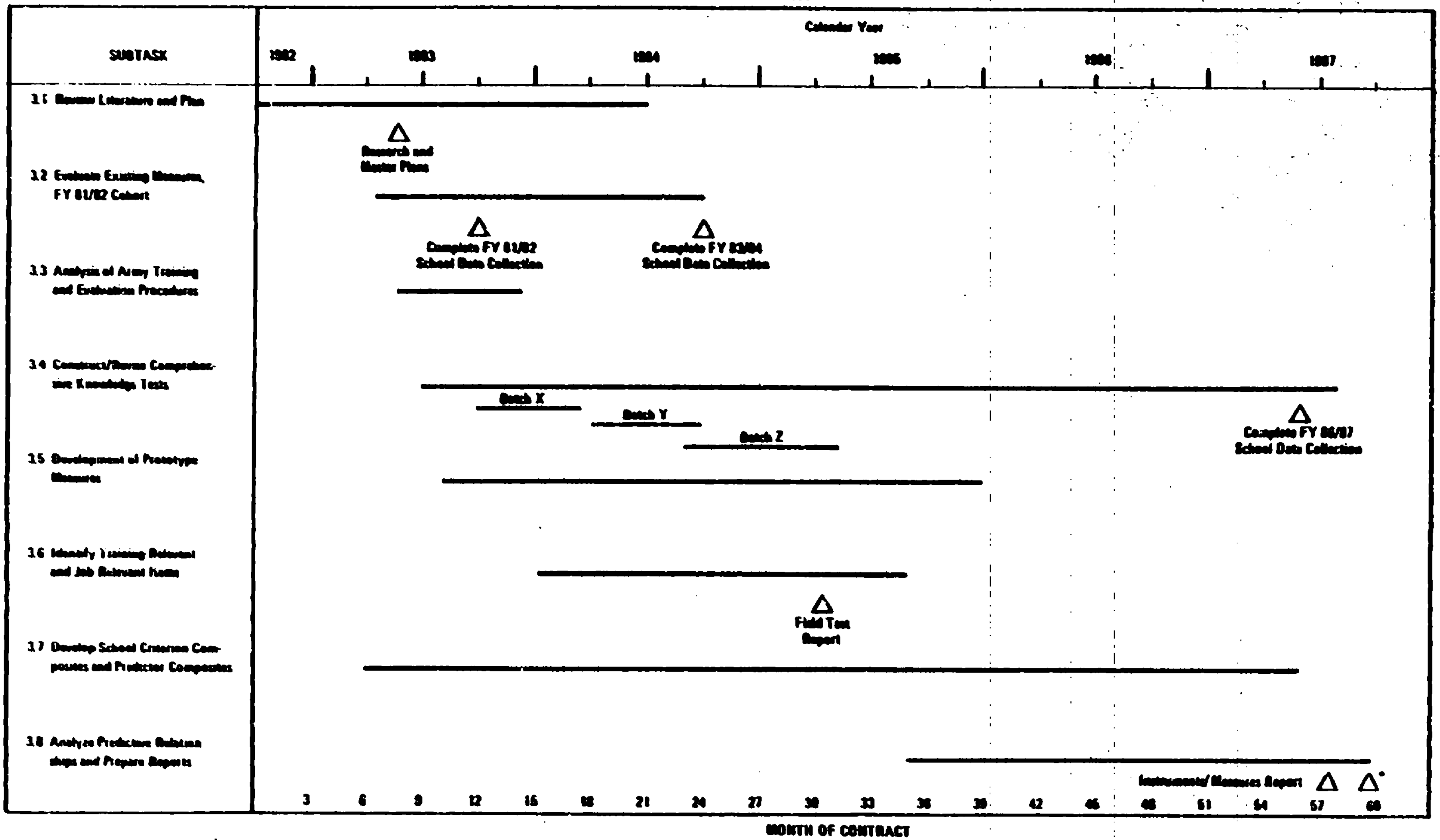
	SME		Trainees/Incumbents	
	# per MOS	days per person	# per MOS	days per person
Subtask 3.2 - Evaluate existing measures.				
Determine congruence of training content measures in school interviews.	3	.5		
Identify 400 trainee records in FY81/82 cohort; arrange copying & mailing of records to HumRRO for both FY81/82 and 83/84 cohort	2	.5		
Subtask 3.3 - Analysis of Army Training & Evaluation Procedures				
Interview training developers & instructors.	3	.5		
Subtask 3.4 - Construct/Administer comprehensive knowledge tests.				
Estimate error frequency in task elements.	3	.5		
Specify elements requiring generalization of knowledge.	3	.5		
Sort elements according to commonality.	3	.5		
Estimate perceptual-motor skill requirements.	3	.5		
Analyze knowledge requirements for elements; empirical testing of trainers to resolve difference.	3	.5		
Try out knowledge tests.			100	.25
Subtask 3.5 - Development of prototype measures.				
Try out free response items and synthetic tests.			100	.25
Develop measures of general performance in training.	4	1		

SME AND TEST SUBJECT SUPPORT REQUIREMENTS
Continued

	SME		Trainees/Incumbents	
	# per MOS	days per person	# per MOS	days per person
Subtask 3.6 - Identify training-relevant & job-relevant items.				
Administer items to entering trainees.			100	.25
Obtain test scores for graduating trainees.			100	.25
Administer items to job incumbents in field tests.			100	.5
Sort knowledge elements according to SME judgment of training relevance.	3	.5		
Subtask 3.7 - Develop School Criterion & Predictor composites.				
Select integrated set of criterion measures.	10**	.5		
** Field Grade Officers.				
Subtask 3.8 Analyze predictive relationships & prepare reports.				
Obtain school data for FY86/87 cohort.			2200 on average	NA***
*** Not applicable since measures will have been operationally implemented in school setting.				

Figure 3

Task 3 Schedule



* Final Technical Report

PROCEDURES

Subtask 3.1: Review Literature and Plan

Except for the construction of conventional tests of achievement, methodologies and concepts appropriate to the analysis, development and application of measurement in training and work situations are not highly developed. Methods for analyzing task and knowledge requirements for both training and job performance, methods of sampling from training and job domains, and procedures for classifying performance requirements and relating them to test formats suitable for producing valid measurement are some of the areas where procedures are not well defined. To benefit from the most recent work and conceptualization in such areas the measurement, educational, industrial, psychological and military research literature will be reviewed. Data bases such as ERIC, NTIS and RDIS will be examined and the following libraries will be included in the search: HumRRO, CTB-McGraw Hill, Navy Postgraduate School, and University of California. Recent work that has not yet been published and reports in the publication process will be sought by personal communication with persons at the various military human research laboratories and other government and non-government research organizations.

Major topics to be included in the review are:

- Job, task and knowledge analysis
- Test sampling
- Behavior classification
- Achievement, performance, and work sample test development
- Performance rating development

Simulation and synthetic testing

Aptitude-performance relationships

Knowledge test-performance test relationships

Training-job performance relationships

Individual differences in training and job performance

Test bias

This subtask will also include preparation of the Task 3 draft research plan and draft master plan. Following ARI review of the draft plans, the revised plans will be submitted.

Subtask 3.2: Evaluation of Existing Measures

The primary purpose of this subtask is to determine whether recent and current training measures can serve as (a) criteria for prior selection and classification measures, and (b) predictors of subsequent job performance.

First, an examination will be made of current measures in the ARI, the Enlisted Master File (EMF), and the TREDS files. Second, visits will be made to selected schools to review the measuring instruments used to obtain these scores, review any additional measures formally recorded by schools, and identify measures not formally recorded but temporarily retained. Scores from measures deemed adequate will be added to the records of the FY81/FY82 cohort in the LRDB and evaluated as criteria for such initial predictors as ASVAB. The scores from these measures will also be added to the LRDB for the FY83/84 cohort where they can be evaluated both as criteria and as predictors.

Evaluation of FY81/82 cohort data and selection of MOS. A compilation will be made with the aid of Task 1 staff of all school performance measures available in the ARI, the EMF, and TREDS, for the selected MOS. The LRDB shows the date of enrollment in the course and whether an MOS was awarded and at what skill level (presumably Skill Level 1) and, if not, the reason for attrition, the disposition of the student (such as "recycled" or "early graduation"), and its effective date, as well as course grade and class rank. What relevant information is available in the EMF and TREDS has not yet been determined.

Collection of data for qualitative analysis. Among the MOS that have been selected for initial study in Project A are:

MOS	Title	Training Site
05C	Radio TT Operator	Ft. Gordon, GA
63B	Vehicle & Generator Mechanic	Ft. Dix, NJ Ft. Leonard Wood, MO
71L	Administrative Specialist	Ft. Jackson, SC
95B	Military Police	Ft. McClellan, AL

Each of these posts is the training site for several of the larger MOS in the FY81/82 cohort. Three other posts, Ft. Bliss, TX, Ft. Sill, OK, and Ft. Sam Houston, TX, are also training sites for a number of the larger MOS and are close enough to the first five to minimize travel costs. We plan, therefore, to visit at least these eight posts to collect qualititative information about the training measures that generated the FY81/82 cohort data.

Using an ARI-furnished printout, the courses taught at these eight posts that are included in the FY81/82 cohort have been identified, and the number of records in the data base were tabulated. The number of records range from 1,672 for the 91B Medical Specialist course at Ft. Sam Houston to 10 for the 32H Fixed Station Radio Repairer course at Ft. Gordon, GA. We expect shrinkage in the data base as the Accessions File, TREDS, and EMF are merged with the LRDB. In addition, some partitioning of the available records will occur since only a majority, not all, of the ASVAB scores in the Accessions File are based on the equivalent forms 8 or 9 or 10. (Scores based on earlier forms of the ASVAB are not equivalent to the 8-9-10 set and, therefore, for certain analyses the training records for persons with ASVAB from the earlier forms will have to be treated as a separate subset.)

For these reasons, most courses with fewer than 500 records in the LRDB were removed from further consideration. The courses with 500 or more records are listed in Table 3-2. It should be noted that Task 1 has informed us that the files have not yet been cleared of duplicate data, i.e., one subject's data repeated several times. Once the file has been checked and cleared of mispunched data, the number of records within each MOS seem likely to change.

Several additional courses are currently under consideration even though the number of records in the LRDB is less than 500. These are courses that have been identified by ARI as having a good distribution on the training measures i.e., sufficient variance to suggest that useful differentiation among trainees was made. These courses are located at Ft. Bliss, TX,

Table 3-2

MOS with 500 Records in ARI
Data Base

MOS	Title	Records	Training Site
05B	Radio Operator	797	Ft. Gordon, GA
05C	Radio Teletype Operator	1046	Ft. Gordon, GA
13E	Cannon Fire Direction Specialist	1164	Ft. Sill, OK
13F	Fire Support Specialist	1133	Ft. Sill, OK
15D	LANCE Crewmember	657	Ft. Sill, OK
15E	PERSHING Missile Crewmember	638	Ft. Sill, OK
16B	HERCULES Missile Crewmember	607	Ft. Bliss, TX
16R	ADS Short Range Gunnery Crewman	939	Ft. Bliss, TX
16S	MANPADS Crewman	887	Ft. Bliss, TX
31M	Multichannel Communications Equipment Operator	1301	Ft. Gordon, GA
31V	Tactical Communications Systems Operator	1133	Ft. Sill, OK
32D	Station Technical Controller	513	Ft. Gordon, GA
36C	Wire System Installer/Operator	590	Ft. Gordon, GA
36K	Tactical Wire Operations Specialist	1304	Ft. Gordon, GA
51B	Carpentry and Masonary Specialist	588	Ft. Leonard Wood, MO
51R	Electrician	573	Ft. Leonard Wood, MO
54E	NBC Specialist	573	Ft. McClellan, AL
62B	Construction Equipment Repairer	1155	Ft. Leonard Wood, MO
62E	Heavy Construction Equipment Operator	1081	Ft. Leonard Wood, MO
62F	Lifting and Loading Equipment Operator	1018	Ft. Leonard Wood, MO
63B	Light Weight Vehicle/Power Generation Mechanic	990	Ft. Dix, NJ
63B	Light Weight Vehicle/Power Generation Mechanic	1680	FT. Jackson, SC
72E	Telecommunications Center Operator	962	Ft. Gordon, GA
75D	Personnel Records Specialist	606	Ft. Jackson, SC
75E	Personnel Actions Specialist	950	Ft. Jackson, SC
76Y	Unit Supply Specialist	1262	Ft. Jackson, SC
82C	Field Artillery Surveyor	927	Ft. Sill, OK
91B	Medical Specialist	1672	Ft. Sam Houston, TX
91C	Patient Care Specialist	1204	Ft. Sam Houston, TX
91E	Dental Specialist	553	Ft. Sam Houston, TX
94B	Food Service Specialist	535	Ft. Dix, NJ
94B	Food Service Specialist	1298	Ft. Jackson, SC

currently identified as one of the eight locations already proposed, and at Ft. Eustis, VA.

Interview procedures and qualitative analysis. Visits will be made to the training sites to examine the measures reported in the computerized files and to identify any other school measures that might predict school success or subsequent job performance or serve as criteria for the initial predictors. During our visits to the schools, measures provided for qualitative analysis will first be classified into paper-and-pencil, hands-on (performance) and other measures (instructor ratings of training performance, number of class hours/days needed to complete course, number of times recycled, etc.).

The number of measures falling in the first or even the second category may be quite large. For example, weekly or even daily spot quizzes (paper-and-pencil) or spot checks (hands-on) might be given in some courses. At a minimum we expect to find a measure following each training module or objective, depending upon the course. When large numbers of measures are taken, we may have to sample them, taking, for example, the first daily quiz and every fifth quiz thereafter. A similar sampling plan can be developed for other measures available in the school records or temporarily held by instructors. Prior coordination with the schools will facilitate development of such a plan.

For each course we will examine:

(1) A syllabus with course objectives, course schedule, lesson outlines.

(2) Copies or descriptions of current measures, including instructions for administration and scoring.

(3) Test scores of the FY81/82 cohort.

(4) Any known statistical properties of the test measures.

(5) Copies of selected individual records from each of a number of courses.

The following questions are representative of the kinds of information to be derived in interviews with instructors and training managers.

A. Characteristics of Measures

1. What training objectives or sub-objectives is this test intended to measure? What portion of the course does this test cover?

2. Why was the particular format (pencil-and-paper, hands-on) chosen?

3. Do the individual items match the elements of training content within the objective or sub-objectives?

4. How were the items generated?

5. Is there a pool of items? How were the items sampled from the pool?

6. Is the item sequence reasonably ordered, e.g., one that reflects the normal sequence of performance?

7. What item format is used?

   a. True-false
   b. Matching
   c. Multiple-choice
   d. Ordering sequence
   e. Identifying right (or wrong) procedures
   f. Open ended response or completion
   g. Rank ordering importance
   h. A mix of formats

8. Which of the following kinds of behavior are required to master the objective?

   a. Decision-making
   b. Application of rules
   c. Selection of strategies

d. Troubleshooting
e. Problem solving
f. Sustained vigilance
g. Immediate (automatized) response to prevent injury to personnel or damage to equipment
h. Attention to fine detail
i. Motor skill
j. Perceptual-motor skill
k. Speeded response
l. Unusual strength
m. Unusual endurance
n. Other

B. <u>Administration of Measures</u>

1. Is the test open book or closed book?

2. Is the test closely or loosely proctored?

3. Is the administration standardized?

   a. Is testing carried out at a central facility?
   b. How are test administrators trained?
   c. Are the instructions given ad lib or read by the administrator?

4. Are equivalent forms available?

5. What are the procedures to keep the test secure?

6. Are questions permitted during the test?

C. <u>Scoring of Measures</u>

1. Is there any subjectivity in scoring?

2. Is the decision to Pass-Fail on the objective covered by the test

   a. criterion referenced (pre-set criterion, absolute), or
   b. norm referenced?

3. Is testing time part of the measure? Do some finish or all finish? If only some, what is the proportion finishing?

4. What is the contribution of the measure to pass-fail on the course?

5. How is the pass-fail decision made at the end of the course?

D. Statistical Properties of Measures, if known

1. Reliability (stability).
2. Internal consistency.
3. Sensitivity to training.
4. Mean differences between racial or gender groups. Differences, if any, to be analyzed to determine source and provide basis for corrective action.
5. Validity.

Similar questions will govern the analysis of the hands-on measures, with certain additions. For example, (a) Is the performance process- or product-scored? What were the reasons for the choice? (b) Is cuing allowed? What are the rules? (c) Is role-playing on the part of someone else required to carry out a particular performance? Who plays the role and how is this person trained? What artificialities occur as a consequence of cuing? As a result of modified task boundaries?

Examples of criteria for the qualitative assessment of measures are given below.

1. The congruence between test items and training content for a specific objective. The number of teaching points in the lesson outlines approximates the required coverage of a measure. When each test item is classified according to the teaching point it represents, redundancies, contaminants and deficiencies in the measure can be identified.

   An index of congruence can be defined as:

$$\frac{\text{Total Number of Test Items} - \text{Number of Redundant Items} - \text{Number of Contaminants}}{\text{Number of Teaching Points}}$$

For example, the test for a module of a course covered by 50 teaching points contains 45 items, twelve of which are redundant and 3 of which are irrelevant to the module (contaminants). The index of congruence is then: $\frac{45-12-3}{50} = .60$

(The appropriateness of a given level of congruence will depend on the type of content being assessed.)

2. Test administration procedures. Points will be assigned for "favorable" responses to the questions on test administration (B 1-6 above) and a total score generated that indicates the relative adequacy of a procedure. Whether a response is "favorable" or not will depend upon the characteristics of the test. For example, a closed-book test would ordinarily be given a point, unless it happens to cover a procedure that could be carried out on the job with the aid of a TM or some other job aid.

The remaining qualitative analyses will proceed along similar lines. The criteria and methods of assigning scores to the various descriptors will no doubt be modified and improved after reviewing actual school measures during the pilot administration of the interview procedures. The purpose

of these qualitative analyses is to identify measures that are likely to have sufficient reliability and validity to warrant being added to the FY81/82 cohort data and to be collected from the FY83/84 cohort, as well as to identify measures that could be improved to increase their reliability and validity. In addition, some of the information obtained in the qualitative analyses will be provided to Task 2 for use in their analysis of predictor and criterion constructs. Descriptions of types of behavior believed to be required to master training objectives, as identified in interviews with instructors and training managers will be particularly useful in the Technical Review conducted by Task 2 staff aimed at identifying the best predictor set. The pertinent training behaviors will be defined, classified by objective and training situation (e.g., lock-step, self-paced instruction) and provided to Task 2 with description of the data collection procedure.

The outcomes of the qualitative analyses will determine what new measures will be added to the LRDB for quantitative analysis. The outcomes may also cast doubt on some measures currently in the data base. For example, a time-to-course-completion measure may not have included extra or after-hours study time. Or the administration of an end-of-course comprehensive test (EOCCT) may be so poorly standardized as to vitiate the scores.

<u>Collection of records for quantitative analysis for FY81/82 cohort</u>. Since the costs associated with reproduction, editing, and entering all available records may be prohibitive in some MOS, printouts will first be obtained of the individual records of the enlisted personnel currently on file in the LRDB for the MOS selected. From the printouts a sample of up to 400 will

then be selected for each MOS, taking into account the date of ASVAB testing, sex, and ethnicity. The school records of these 400 will then be identified and copied. Should initial contact with the schools make it apparent that such an approach is not feasible (that is, if searching for specific records is too cumbersome) 400 records will be sampled sequentially from all 81/82 classes and sorted later for date of ASVAB, sex, and ethnicity.

Collection of records for FY83/84 cohort. Currently there are no data in the LRDB for the FY83/84 cohort. The measures determined by the qualitative analyses to be most promising will be input to the LRDB. Arrangements will be made with schools to send us records on a continuing basis as each new class completes training, starting in July 83. Records for every trainee entering a course will be requested to increase the likelihood that the number of records is sufficient to conduct the follow-up data collections during the first and second tours when considerable attrition can be expected. Schools will be requested to include additional measures in school records that appear promising on the basis of the qualitative analysis but which were not included in the data collected by ARI for the FY81/82 cohort. Also, emphasis will be given to measures which appeared qualitatively acceptable for the FY81/82 cohort but which were found so incomplete in the data base that no quantitative analyses were done.

In addition, if any of the 19 MOS selected for Project A are not in the group of MOS selected for the FY81/82 cohort, the measures for such MOS will be assessed qualitatively and quantitatively. That is, the analyses performed for the FY81/82 cohort measures will be repeated for any MOS

not in our FY81/82 set. Just as before, this will entail visits to schools to become knowledgeable about their measures and to obtain records of trainee scores.

Quantitative Analysis. The quantitative analysis of the existing measures will focus on the adequacy of the distributions of scores from these measures to support the statistical analysis projected for them. Clearly, the number of scores available is one important criterion for deciding whether a set of scores is a suitable basis for validation and prediction research; the extent to which the measures appear to discriminate among the trainees is clearly another, but there are no convenient rules of thumb for accepting or rejecting a distribution on either criterion. It is necessary to devise some scheme whereby distributions of scores currently available and to be encountered in the future can be classified according to their potential utility for further analyses.

To this end, a set of some 25 distributions of scores will be selected out of the data from the FY81/82 cohort (once those data have been purged of redundancies, drops, and recycles) and out of the data from other measures currently in use. Such indices of central tendency, variability, skewness, and kurtosis as the mean, median, mode, minimum and maximum scores and their standard score equivalents, score range, score variance, standard deviation, semi-interquaretile range, coefficient of variation, Pearson's second coefficient of skewness, a measure of kurtosis, the quartile scores and the scores at +/- 1 standard deviation will be determined for each distribution. In addition, a histogram will be made of each distribution. A total of 15 to 20 experts in test construction and statistical analysis

will be selected from the staff of Project A, and asked to rate each of these distributions and their associated statistics on a five-point scale as very useful, useful, marginally useful, doubtfully useful, or not useful in the statistical analyses involved in validation and prediction.

Once the distributions have been rated by the experts, analysis will show what characteristics of a distribution can conveniently be used for accepting or rejecting it as a basis for further analyses, and these characteristics will be used as the basis for accepting or rejecting existing measures to be retained in the LRDB.

As FY83/84 cohort data are input into the LRDB, descriptive statistics (means, variances, scatter plots, etc.) and correlation matrices will be computed to determine whether some measures should be dropped from further consideration. The descriptive statistics will be used to screen such obviously inadequate measures as those with no variance.

We will stay in touch with the schools to identify changes made in lesson outlines and school measures during the data collection period for the FY83/84 cohort. Changes in course outlines or training measures, such as adding or deleting blocks of material, or altering the match between training and test content, could seriously affect the comparability of measures between classes. (This will be a continuing problem within and between cohorts.)

Certain of the school measures that were input to the LRDB originally from the ARI files or added during this study will no longer be of interest. These will include: (a) measures with intractable shortcomings uncovered

in qualitative analysis, e.g., measures that appear impossible to standardize; (b) measures that fail to differentiate among individuals, and (c) measures that entered the correlation matrix but appeared to be minimally related to anything else.

On the basis of the qualitative and quantitative analyses of existing measures, we will prepare a report listing our recommendations for improving the administration and scoring of measures in each of the 19 MOS in Project A. This report will be forwarded to the school commandant and the director of training at each site, upon approval of the ARI COR.

With the approval of the director of training, we will brief each recommendation to the instructors and course chief involved, and describe specifically each modification that we would like to see implemented. In particular, our briefing to the instructors will explain the rationale for each change in terms of expected improvement in reliable and valid information to be obtained about each trainee's level of mastery of course objectives.

Subtask 3.3: Analysis of Army Training and Evaluation Procedures

An essential aspect of the development and improvement of school measures is the determination of the relation of these measures to actual job requirements. Even more fundamental is the relation of the content of training and its measures to the requirements of the job itself. The primary purpose of this subtask is to determine the content validity of current training and training measurement by: (a) identifying the

processes by which these components of the training system are derived, and (b) establishing their relation to job content. The general method by which this will be done will be to track the requirements identified by job analysis, through the training development process, to their representation/nonrepresentation on within-course and end-of-course evaluations.

In the ideal training development system, there would appear a simple one-to-one correspondence between job requirements at one end of the development process, and the content of trainee evaluation at the other. Experience with attempts to reach this ideal has demonstrated that such a correspondence is difficult to achieve and not likely to be typical of the real-life training process. In practice, the content of training can arise from several sources in addition to identifiable job requirements (past practices, command preferences, rational analysis, instructor proclivity, tradition, etc.). The focus of this subtask will not be a judgmental comparison between MOS as to the quality of training development efforts, but rather, a determination of how the content of training and the content of trainee evaluation come into being.

Tracking the progress of training development through its various stages requires on site interviews with training developers themselves and examination of training development products. In this subtask, semi-structured interviews will be used to collect information across individual courses. Interviewers will follow a guide, but they will be able to deviate into promising areas of inquiry. By using this method, it is possible to locate the source of training content as well as the eventual destination of training development products (e.g., training objectives,

training activities, training measures). By interviewing persons who are responsible for different parts of the training development process, it is possible to obtain information about the adequacy of the output of one phase for serving as the input to the next phase, and to piece together a picture of the whole process as it actually occurred for a given course.

A recurring problem in tracing the relation of job requirements to training activities and measures should be mentioned. A common component of the training development process is the specification of training objectives as an intermediate step in the derivation of training. We have found (Vineberg & Joyner, 1980) that while the specification of training objectives is virtually universal, the procedures used to identify objectives are highly variable and frequently unclear. There is evidence that objectives are often prepared after the fact and are derived from training content rather than used to generate it. Where records are maintained, formats for displaying the relation between tasks and training objectives often make it hard to determine what objectives have been derived from a given task. That is, tasks are often listed by objective, rather than objectives by task.

Accordingly, we plan to "map" the derivation of training development products--e.g., training requirements, training objectives, training activities, and training measure content--within a matrix that cross-references these products by job task requirements. To accomplish this mapping, the analysis will treat the various aspects of the training development process independently.

Job Requirements Analysis. Personnel who know about the development of the course will be interviewed, and job/task analysis data examined, to determine the following:

- o how job requirements were identified or defined;
- o how training requirements were identified or defined;
- o how training requirements were derived from job requirements; and
- o how training requirements were specified/transmitted to training designers/developers.

Of special concern at this point is the question of what happens to job/ task analysis information after it has been developed.

Instruction Developers. Personnel who design and develop training will be interviewed, and records of training design decisions will be examined to determine: (a) how job/task analysis information is put to use; (b) how the content of lesson plans, texts, and other media of instruction is selected; (c) what formal specifications are given to training developers, etc. (Training developers may be the trainers themselves, or other personnel. Even when schools are organized into separate divisions to carry out different training development functions, considerable functional overlap has been found.) In this subtask, questions will concentrate on how instruction content is derived, developed, and evaluated rather than how decisions about instruction technique are made, e.g., media, methods, self-paced vs. group-paced, even though information about the latter will undoubtedly emerge in the course of interviews.

Test Developers. Developers of evaluation instruments will be interviewed, and evaluation instruments examined, to determine how the content of trainee evaluation is selected. Of special concern at this point is how the domain to be sampled during criterion measurement has been determined. Are tests based on training content (as represented in texts, lesson plans, and other media), job/task analysis information, or some other source? Are tests constructed by the instructors who conduct training, or by other personnel?

Trainers' descriptions of the output from and relationships among components of the training development process will be confirmed by examining lesson plans, texts, tests and other developmental products. Each job task selected for training will be classified as represented or not represented in training and training measurement.

The information collected in this subtask will be used subsequently to interpret predictive relationships that are obtained between measures of training performance and measures of MOS-specific performance. In addition, the information collected will be used in the development of the knowledge tests and prototype measures (Subtasks 3.4 and 3.5).

Subtask 3.4: Construction/Revision of New Comprehensive Knowledge Tests

Improved knowledge tests to serve as criterion measures of school performance and to provide a basis for analyzing the mechanism by which school performance is predictive of subsequent job performance will be developed in each of the selected 19 MOS. Test items will be derived from

requirements of tasks specified in Soldier's Manuals for which task performance analysis (TPA) information is available. The knowledge tests will contain subsets of items reflecting training requirements and job requirements in situations where it is appropriate to make this distinction. In some MOS the content of training and the requirements of a job will be more or less identical; in others, information acquired in training may not be retained because it is less relevant to job performance. Where a sufficient number of test items can be developed for both classes of information, we believe it will be possible to provide a more accurate estimate of the role of an individual's capacity to learn versus his or her acquisition of job-specific information, as a predictor of job performance.

Since it will be efficient to use test items that are currently used in Army schools, wherever possible, the plan for test development calls for using existing test items whenever they are content valid (as determined in Subtask 3.2) and psychometrically adequate.

Test item development will begin for a subset of 6 of the 19 MOS (05C, 13B, 19E/K, 63B, 71L, and 95B) immediately after approval of this Research Plan and the first Troop Support Request is received. Test development will occur for two groups of 6 MOS and one of 7 MOS on a staggered basis during the approximate period October 1983 - December 1985. The major steps in the construction of improved test items and their assembly into end-of-course comprehensive tests are as follows:

1. Obtain TPA information for each task in the Soldier's Manual for each selected MOS. TPA information will be obtained from the analyses conducted by RCA (Contract NO. DABT 60-81-C-0017) and supplemented by other task analysis, information, as required, from Army schools, from SQT notices and tests from Task 5, from job publications such as field manuals and technical manuals, and from analyses performed to develop Training Extension Course (TEC) lessons. Preliminary inspection of a small sample of the RCA TPA is encouraging, and they are expected to be the most reliable source of task information for the MOS not treated in Task 5.

2. Subject TPA information for each task to the following analyses:

   a. Three SME independently estimate frequency of errors in performance (5-point scale) for each behavioral element of TPA.

   b. Three SME independently designate elements of behavior in TPA requiring application of a rule or principle and thus involving generalization of information - e.g., how to place antenna for best reception. Discrepancies resolved by research staff.

   c. Research staff sort elements of TPA into common and non-common behaviors across tasks to identify generality of elements. Basis of assigning commonality and level of behavioral description in TPA at which it can be specified to be determined when TPA data become available.

   d. Two SME-research staff teams independently judge presence or absence of application of perceptual and/or motor skill in performance of each behavioral element of TPA. Those elements judged to contain these requirements become candidates for hands-on or prototype measures (see Subtask 3.5). In those instances described above when perceptual motor behavior calls for using a hands-on test, conventional knowledge test items will also be developed to permit examination of the relationships between hands-on and paper-and-pencil test measurement of the same job performance component.

3. On the basis of information generated in Step 2, and for the 25 percent of TPA elements with highest error ratings, compute the proportion of procedural elements to total elements (elements involving application of rules/principles [generalization of information] plus number of procedural elements). Repeat the computation for the 25 percent of the elements with the greatest commonality. Compute the mean proprotion of procedure application to total elements for the two domains. This average proportion provides a basis for determining the selection ratio of procedure to rule/principle items in the tests. This procedure has been devised on the supposition that procedural elements may account for the bulk of performance requirements while errors may be disproportionately associated with rule/principle application elements. Thus the procedure employed will cause the proportion of procedure and rule/principle application represented on the test to reflect the proportion of these two domains in the more error prone parts of performance.

4. Standardize error and commonality scores and compute combined scores for all elements. Rank order all procedure application elements and all rule/principle application elements separately in descending order of combined error/commonality score. Select elements with highest combined scores in each domain for representation as knowledge test items in accordance with the selection proportion determined in Step 3.

5. Analyze and specify knowledge requirements for each performance element selected for test item construction. Statements of knowledge required for performance of each selected element will be prepared independently

by three SME. These statements will be compared by a member of the research staff who will attempt to resolve differences.

6. Construct knowledge test items or identify comparable items currently in existing measures at Army schools. Statements of knowledge requirements generated in Step 5 will be translated into knowledge test items, in accordance with standard prescriptions for the construction of knowledge test items (e.g., Adkins, 1947; Anastasi, 1976). As indicated above, two major types of knowledge are to be represented in the tests: knowledge that must be generalized in its application and knowledge that does not. A crucial aspect of test item construction, therefore, will be to insure, insofar as possible, that performance elements of tasks that involve the generalization of knowledge maintain this requirement during translation into test items. In all instances where the generalization of knowledge is called for, we will endeavor to generate items for this category of performance that require the application of rules or principles rather than the recognition, recall, or restatement of rules or principles.

The total number of test items to be developed will vary as a function of the number of different performance elements in an MOS. Although ultimately the length of knowledge tests will be constrained by the amount of time available in training for test administration, these constraints will not apply during the development phase. It is expected that a fairly large number of items (300-400) will be developed in order to increase the likelihood of ultimately capturing a significant number of both training-relevant and job-relevant items.

7. Item tryout and revision. Test items will be administered to two groups of 50 trainees as they are developed to assure clarity, comprehensibility, meaningfulness, and relevance of content to performance. These characteristics will be examined in post-test discussions with the test subjects. During the interval between the two tryout administrations, changes indicated by the results of the first administration will be made in the tests.

Test items will have been constructed and considered content-valid on the basis of their representation of a sample of common and error-prone elements of task performance on an MOS. Items will not have been selected, for example, on the basis of their capacity to discriminate among students in training. In Subtask 3.6, however, the comprehensive knowledge tests will be administered to persons prior to the beginning of training, to persons at the end of training, and to persons who are performing as job incumbents. Items that reveal little capacity to detect knowledge acquired either in training or on the job will be discarded subsequent to the field test in which job incumbents are tested.

After trial administration to trainees and any necessary revision, test items will constitute a pool from which at least two alternate test forms can be assembled. The alternate test forms will be developed by random sampling, stratified by commonality and error values.

8. The job knowledge tests will be given to 100 trainees at the <u>beginning</u> and <u>end</u> of their AIT courses and to an average of 500 job incumbents in each of the MOS along with the other criterion measures developed for administration to the FY83/84 cohort. Following analysis of the data and possible revisions of the tests, arrangements will be made for the continuous administration of the tests to the FY86/87 cohort at their respective schools. These tests will provide a principal basis for establishing predictive relationships between school performance and measures of both hands-on job performance and Army-wide performance.

<u>Subtask 3.5: Development of Prototype Measures</u>: Evaluation of Measures of Free Response, Synthetic Testing of Hands-on Performance, Measures of General Performance, and New Performance Indices

This subtask focuses on research designed to develop the means for evaluating elements of task performance and other aspects of training performance not amenable to evaluation with conventional knowledge tests. New methods of testing and measurement are desirable in at least the following areas: evaluation of elements of task performance in situations where a free response is necessary to avoid the artificial cueing of a response; evaluation of task elements that involve perceptual-motor skill; and evaluation of indicators of general performance and effectiveness in training not related directly to the performance of MOS-specific tasks. In addition, it is desirable to develop new indices derived from existing tests and measures to provide information not available in the measures themselves.

The overall objective of this subtask is to develop an array of prototype measures that might profitably be tried out as indicators of training success, criteria for selection research, or predictors of later job performance. The subtask is not meant to yield a fully developed and validated set of measures that will then be ready for implementation as operational measures of training achievement or final versions of criterion measures for research. Consequently, the item tryouts and test development work for free response and hands-on tests will be limited to a relatively small number of MOS (the four MOS that will receive the Trial Predictor Battery--05C, 19E/K, 63B and 71L). However, the measures of general performance and the new performance indices will be developed for all 19 MOS.

General Procedures. In general, the procedure will be to select test content for each of the prototypes by modifying existing measures that have been developed or are used in some other context. Once an array of potential prototypes has been identified, the individual candidates will be described comprehensively by Project A staff. The candidate measures, will be then subjected to expert review, much as the potential predictors are being reviewed in Task 2. That is, they will be evaluated in terms of their strengths and weaknesses by panels of research psychologists.

Prototype measures that survive the review process will then enter along with the knowledge test items the steps of item development, content validation by SME, and pretesting on tryout samples.

1. Free response measures. The traditional use of multiple choice tests with identified response options is not adequate for all forms of knowledge testing. In many instances it is desirable to evaluate the knowledge that mediates performance of a particular element of a task without any of the artificialities implicit in multiple choice items. A test of rule application, for example, should be constructed to provide opportunities to reveal awareness of the relevance of the rule to the task or situation at hand as well as mere knowledge of the rule itself. It is difficult with a multiple-choice test to define trouble shooting, problem-solving or decision making tasks while simultaneously maintaining the conditions of uncertainty that are characteristic, if not the essence, of these types of situations.

Although such considerations as these might imply that hands-on testing is the only viable format, there are, of course, a number of methodological and practical difficulties to be considered in utilizing hands-on tests in school settings: (a) variation can be expected in administration and scoring; (b) the number of tasks that can be evaluated is limited by time constraints; and (c) hands-on tests are costly and time consuming to construct as well as administer.

The general approach taken in Task 3 to the measurement of task performance described earlier (Subtask 3.4) is to decompose tasks into their elements and sample among these elements for purposes of assessment. With such a strategy it seems possible to allow a free response format for a limited number of task elements in any given measurement effort.

We will examine the feasibility of prototype measures of three types of free response: (a) open-ended written items for group administration, (b) verbal response in a one-on-one testing situation, and (c) performance/demonstration in a one-on-one testing situation. These three formats are likely to vary both in efficiency of administration and suitability according to the property of the task elements being measured and the characteristics of the population being tested. Thus open-ended written items are clearly most efficient but only may be suitable for use with persons with above average verbal ability or in situations where the correct response is commonly described in one or two words. For persons of lower verbal aptitude, performance or demonstration of the task element may be most appropriate.

We will examine the feasibility of free response measurement of task elements in four MOS for which hands-on performance measures have been developed (Task 5) and to which the Trial Predictor Battery will be administered so that their relation to these measures can be determined. The following general procedure will be used:

(1) Prepare detailed research and development plans for development of free response prototype measures. Prepare troop support request. Submit plans and troop support request to ARI.

(2) Via judgments of SME, designate elements of task behavior requiring application of rules or principles or which otherwise are suitable candidates for evaluation via free response measurement.

(3) Construct a pool of free test evaluation items in three formats: open ended written, verbal, and performance/demonstration.

(4) Try out and revise items on two trial samples of 50 AIT trainees in each MOS for which a free response item

pool was constructed. Evaluate items in terms of difficulty, understandibility, and feasibility of standardization for operational use.

(5) Administer items to sample of 100 trainees at beginning and end of AIT course (see Subtask 3.6).

(6) Administer free response items during the administration of the FY83/84 cohort first-tour performance measures (which include the hands-on tests for tasks from which the free response items were derived).

(7) Analyze predictability of hands-on task performance by free response task element format, trainee characteristics, and task element characteristics. Report results.

2. <u>Synthetic testing of hands-on performance</u>. Hands-on tests are generally recognized as the method of performance evaluation that provides the most direct and complete means of assessing task proficiency. The major need for hands-on tests occurs, however, in tasks that call for a display of perceptual-motor skill. Such skill cannot be represented adequately in a conventional paper-and-pencil test of job knowledge. Unfortunately hands-on tests which are appropriate for such measurement are expensive in cost and time to develop and administer. For obvious reasons, psychologists have frequently sought measures represented in various kinds of simulations as substitutes for a full hands-on test, i.e., display of criterion behavior in a realistic criterion setting. One such attempt has been that of Osborn and Ford (1976) who coined the term synthetic testing to refer to "... a job performance test that has been degraded to some degree in the range of task elements covered or in the fidelity of stimulus/response features".

Osborn and his colleagues have explored a variety of issues in synthetic testing, but the basic strategy involves determining whether a task can be

meaningfully partitioned into subtasks or steps, partitioning it into those parts and then, depending on the characteristics of the task and purpose of testing, following one of these procedures:

- Select a test modality for a part-task test of the most difficult element.
- Test all task elements using a method appropriate for the most difficult one.
- Test the most difficult task element with its most appropriate method and the remaining parts of the task with the most efficient method.

In the present examination of prototype synthetic testing a variant of the first procedure will be used: hands-on tests will be designed for task elements (two) that call for a display of perceptual-motor skill. The predictability of task performance from synthetic test performance will be examined in the four MOS using the following general procedure:

(1) Prepare detailed research and development plans for development of synthetic test prototype measures. Prepare troop support request. Submit plans and troop support request to ARI.

(2) Identify elements of task behavior requiring display of perceptual motor skill. (See procedures Subtask 3.4.)

(3) Construct synthetic hands-on prototypes for perceptual motor task elements using the methods outlined in Osborn and Ford (1976). The initial test content will be derived from hands-on items developed in Task 5 and from existing (if any) hands-on measures currently in use in AIT schools. The items will be revised and "degraded" by Task 3 staff in consultation with Task 5 staff.

(4) Try out and revise prototype items on the two trial samples of 50 AIT graduates (along with the job knowledge and free response measures).

(5) Administer measures to sample of 100 trainees at beginning and end of AIT course (see Subtask 3.6).

(6) Administer revised synthetic tests and Task 5 hands-on performance tests for tasks containing perceptual-motor elements to the FY83/84 cohort.

(7) Analyze predictability of hands-on performance from synthetic test performance by trainee characteristics. Report results.

3. Measures of general performance in training. The model of soldier effectiveness developed by Task 4 includes two types of measures: objective measures, (attrition, number of AWOL, awards given, etc.), and performance ratings. In the development of measures of general performance in training, we will perform the following steps. We will develop a separate checklist of questions for each of the two types of measures. These checklists will provide information as to both the appropriateness of the dimension for the training environment and the feasibility of measuring the dimension in a school setting. For example, typical questions asked about each objective measure, such as absenteeism, would be:

(1) Is this measure recorded accurately?

(2) Does this measure tell you anything about a trainee's performance in school?

(3) Is this measure currently collected by the school?

(4) What other measures should be collected that relate to the performance dimensions?

(5) Is there much difference in trainees on this measure?

In addition to the above, typical questions asked about ratings would include:

(1) Can you apply this rating method in your class?

(2) Do you have time to do so?

(3) Do you have sufficient contact to rate trainees?

(4) Could we change this measure to fit your needs? How?

(5) Do you feel this particular dimension or scale is relevant to job performance?

(6) Would you support the development or use of such rating scales in training?

(7) Are there any rating scales currently being used?

(8) Are there other ratings that should be developed to measure training performance?

We will administer each checklist to 2 course instructors, and 2 Army unit leaders in each of the 19 MOS. We will administer the checklists to the Army unit leaders because it is expected that school personnel might not be familiar with the value of some of the objective measures kept at the unit level.

The basic approach to development of these prototypes will be to begin with the objective indices and Army-wide rating scales under development in Task 4 and evaluate the feasibility of using them in school settings. Initially, at least, it seems quite reasonable that rating dimensions such as "overall performance" or "demonstrated commitment to the Army" could be modified to be used by school personnel to rate students.

Once modified, the candidate measures would be evaluated by Project A and ARI staff and by a panel of instructional staff to identify major shortcomings and gaps in the array of scales that could be filled in by additional development work.

If major gaps in the array of dimensions for rating general factors in training performance are identified, e.g. "mastery of factual knowledge", Task 4 procedures will be used to construct new scales that have been so identified.

We will pilot test the measures selected on the same sample of trainees to whom the revised knowledge tests will be administered, and collect descriptive data about the distribution of the scores of these measures. Such measures will include the mean, median, mode, standard deviation, coefficient of variation, interquartile range, etc. The decision as to which of the measures are the "best", i.e., have the most potential for predicting job performance, will be made on the basis of the standards set up by the experts in Subtask 3.2 for existing measures. The most appropriate means and frequency of administration for these measures is dependent on the characteristics of the measures and specific situational requirements.

The measures chosen will be discussed with training managers at schools for each of the 19 MOS where we will seek implementation. Data will then be collected on a continuing basis as with other school measures for the later portion of FY83/84 cohort and if, based on analysis of the FY83/84 concurrent validation data they are apparently effective predictors or criteria, on the FY86/87 cohort.

4. New Performance Indices. Although it is difficult to talk about specifics before our compilation of information from training schools is complete, new indices of performance will be formulated whenever a combina-

tion of existing measures or refinements of these measures offer promise of providing useful information not contained in the original measures themselves and the refinements can be accomplished in an economic and useful fashion. For example, combined measures of learning time and achievement level may generate an index that would be reasonably easy to determine and which would be significantly related to later performance measures.

We will compute such indices and test their relationship to job performance measures obtained from the FY83/84 cohort in the 19 MOS. Unlike tests, such as job knowledge tests, which can be administered out of context to incumbents, these indices can only be validated longitudinally. This is because they include actual training outcomes (such as time spent). This disadvantage is offset by the fact that they are readily obtained from all trainees, and require no additional test administration time.

In the process of obtaining time measures, several factors must be accounted for or controlled, since the computation of training time routinely performed for administrative purposes is influenced by several extraneous factors.

In a self-paced course, if trainees determine when to present themselves for testing, it is ordinarily not possible to determine at which point they reached acceptable mastery (Christal, 1976). One student with a passing score may have learned enough to pass just prior to the test; another, following a more conservative strategy, may actually have reached criterion well before he or she elected to be tested, and then spent additional time over-studying the lesson.

In many courses, adjustments are made for time lost to sick call, non-training extra duty, etc. The method of accounting for this time varies considerably. Closer control over the accounting for these and other factors is expected to be needed more for the purpose of Project A than is currently needed for administrative purposes.

Draft and final plan for prototype measure data collection. The introduction of prototype measurement into on-going training programs, apart from the generation of new indices from existing measures, is dependent on the acceptance of such measurement by course managers and instructors. Because the new measures will be developed with the cooperation and assistance of course personnel and because considerable attention will be devoted to explaining the purpose and benefits of the measurement, it is expected that in general the acceptance of the new measurement methods into on-going training will not present a problem. Where the use of the new measures imposes additional personnel requirements, e.g., test administrator/scorers for hands-on tests, such demands may not be as readily met.

The actual requirements and the ways in which the new measures will be administered cannot be anticipated since they depend, of course, on the particular characteristics of the individual training programs. When development of the new measures has begun at individual schools the Task 3 staff will prepare a draft plan for data collection from trainees in the FY83/84 cohort, and submit the plan to the COR for review. The plan will discuss types of measurement, administration procedures, frequency and locus in training of measurement and scoring procedures. After review by

the COR the draft plan will be revised as necessary and a final version submitted. A similar plan will be submitted for data collection in the FY86/87 cohort.

Subtask 3.6: Identification of Training-Relevant and Job-Relevant Knowledge Test Items

To provide a basis for interpreting predictive relationships between measures of school and job performance, the relevance of knowledge test items to training and job content will be determined. After the knowledge tests have been revised on the basis of the tryouts, they will also be administered to samples of entering trainees and samples of job incumbents in order to assess the effects of training and job learning. The testing of job incumbents will occur as part of the first and second field tests to be conducted during April-June 1984 and November 1984-January 1985, respectively. In these field tests knowledge tests will be administered to incumbents in the 9 MOS for which hands-on performance measures are being constructed in Task 5, MOS in which training/job performance relationships can therefore later be analyzed.[1]

When the test data are available, test item difficulty indices will be compared for entering trainees, graduating trainees, and job incumbents. Items that reveal a decreased difficulty (higher percentage correct) as a consequence of training will be defined as "training-relevant"; items that

---

1Hands-on tests (Task 5) will be administered to incumbents at skill level 1 in 4 and 5 MOS during the first and second field tests, respectively.

reveal little change as a result of training but which reveal a decrease for job incumbents will be defined as "job-relevant"; and items that demonstrate a decrease both at the end of training and among job incumbents will be defined as "training- and job-relevant." It is expected that most items will fall in the third category.

In order to isolate the mechanisms that mediate relationships between training performance and job performance using a cross sectional design, at least two assumptions must be made:

(1) course content in a particular MOS remains relatively constant over time, and

(2) the groups in question are well matched with respect to any variables that may affect job knowledge test scores such as intelligence, race, sex, etc.

To the extent that these assumptions are not met, the meaning attributed to the subset of items identified in Subtask 3 becomes suspect. For example, differences in scores on the knowledge test may be due to group differences in intelligence. If this were the case, the items might be more appropriately labeled "intelligence relevant" than "job-relevant" or "training relevant." Also, differences in scores on the knowledge test may be due to group differences in training content. Thus items that were formerly, but not presently, included in training could be erroneously interpreted as job relevant merely because job incumbents who received an earlier version of training showed higher mean scores. Shifts in training content are not infrequent in military instruction and may represent a more pronounced threat to construct validity than failure to match groups.

To avoid these problems it is planned to use a longitudinal design in which a sample of 100 trainees is tested upon first entering AIT and then retested at the end of their training. A counterbalanced design will be employed with two groups of 50 trainees first taking alternate forms of the knowledge test and later taking the other form. Our estimates (p. 19) indicate between 19% and 52%, depending on MOS, of these soldiers will be among the 500 tested during the FY83/84 cohort performance data collection. If we assume 30 are required for meaningful retest results, then, about 16 MOS will have sufficient data. Consequently, we will be able to compare performance across time with more rigor than could be accomplished with independent groups.

As mentioned earlier, the ratings of general performance in training, the new performance indicies, and, in the case of 4 MOS (05E, 19E/K, 63B and 71L), the free response and synthetic measures, will also be obtained for the sample of 100 trainees. The interrelationships among the new as well as existing school measures will be determined as well as their separate relationships with existing and new predictors. Of particular interest will be any shifts in relationships between predictors and school measures that occur over time e.g., a drop in predictability of knowledge test scores from AIT entry to graduation to job incumbency.

The data from the longitudinal research will be analyzed to determine the amount of variance in job performance explained by the various subsets of items. First, knowledge items will be partitioned into one of four groups. Group 1 will consist of items that do not require generalization and are learned in training. Group 2 will consist of items that do require generalization and are learned in training. Groups 3 and 4 will consist

of items that are learned on the job (are less difficult for incumbents) and are procedure-related or rule-related, respectively.

Several multiple regression equations will be developed and tested within each of the 19 MOS. The equations will consist of a job performance measure (as the criterion), one or more measures of cognitive ability and other pre-induction predictors, and the four groups of test items defined above, as follows: Job Performance = Pre-induction Measures + Group 1 + Group 2 + Group 3 + Group 4.

This model will be tested against a model that includes only the pre-induction measures: Job Performance = Pre-induction Measures.

If there are no significant differences between these two models, as tested by an "F" procedure outlined by Cohen and Cohen (1975), it would appear that the knowledge items are not capturing any unique or additional portion of the criterion variance. If, however, one or more type of item consistently enters into the equations with a significant F value, we will be able to tailor further knowledge test development in the direction of increasing those types of items.

In addition to the empirical methods of identifying training-relevant and job-relevant test items described as above, judgments of trainers will also be used. In each MOS, three SME will independently sort behavioral elements from task analyses into those that should be mastered by graduates of their course, considering the course objectives and content, those that should be mastered on the job, and those that are not job relevant. The

correspondence between the test item content and these behavioral elements will also be judged by the SME. This procedure will provide a means of verifying and interpreting the results of the quantitative analysis.

Subtask 3.7: Develop School Criterion Composites and Predictor Composites

Candidate training measures to be used in validating initial predictors and in predicting MOS-specific and Army-wide performance will be identified or developed beginning in Subtask 3.2 with the review of existing measures and continuing through the development and tryout of prototype measures in Subtask 3.5 and in the further refinement of the job knowledge tests in 3.6.

The relationship of these measures to existing and newly developed predictors will be examined through administrations of predictors to FY83/84 cohort personnel from October 1983 through June 1984. The relation of training measures to subsequent criteria will be examined through administrations of Army-wide and MOS-specific measures to FY83/84 cohort personnel from June through September 1985.

On the basis of these data collections, training measures will be assembled into integrated sets of criteria and predictors. The construction of predictor sets will be essentially empirical. Candidate predictor measures will be factor analyzed, and regression techniques will be used to identify groups of measures that contribute the most to the predictability of the various criteria that will then be available.

Building criterion sets, however, will also require judgments about the relative contribution of different measures to the construct "school success". Since school success cannot validly be defined simply as whatever components of training are predictable, the judgments of school personnel will be obtained to inform this selection and weighting process. For example, in self-paced courses two potential measures are: (a) total time to complete the course, and (b) mean score of first attempt to pass module tests. Which of these is more important, to spend the least amount of time in training or to obtain the highest possible score?

If these two measures are found to be highly correlated (which is not unlikely, since students of higher ability may score high on both counts) there is no serious conflict to be resolved in weighting them in a criterion composite, since alternate sets of positive weights would produce composites that are highly correlated. But if these or other measures are less highly correlated, a value judgment must be sought. We will convene panels of officers for each one of the 19 MOS in Project A for this purpose. Panel participants will be presented with descriptions and explanations of the candidate criteria which the panel will be asked to weight according to their contribution to the definition of a "successful trainee."

These criterion composites will then be used in the longitudinal validation of initial predictors using the FY86/87 cohort. Monitoring of the administration of these measures will take place from March 1986 through February 1987.

Subtask 3.8: Analyze Predictive Relationships and Prepare Reports

As data become available from the administration of performance measures to the FY83/84 cohort (June-September, 1985), correlation matrices will be used to identify groups of measures that are highly related to MOS-specific and Army-wide criteria. This will involve existing school measures reviewed in Subtask 3.2, prototype measures developed in Subtask 3.5, and job-relevant and training-relevant knowledge test items identified in Subtask 3.6. Cluster analysis will first be used to determine which types of training measures are most related in general to the various types of criteria. All analyses will be coordinated with Task 1.

The predictive relationships between the new (Task 2) predictors administered to the FY83/84 cohort during the same period will likewise be determined, using the same methods as used in Subtask 3.2 to determine relationships between the current ASVAB and current school measures. These analyses will be used to help select the integrated sets of training measures to be used in the main longitudinal validation with the FY86/87 cohort, beginning March, 1986.

The relationship of new predictors administered to the FY86/87 cohort to Task 3 training measures will be analyzed as data cohort members pass through training and data on both measures become available (approximately July 1986 through April, 1987).

The results of these analyses will be presented in a draft technical report and draft instruments report which will be revised on the basis of comments

from the COR. The final technical report will be prepared in two parts. The first part will be an executive summary of the types of information useful to Army personnel managers; the second part will be a detailed narrative in a format suitable for ARI publication and submission to psychological journals.

SUMMARY OF EXPECTED OUTCOMES FROM TASK 3

Operational Outcomes

1. Integrated sets of training measures for a sample of MOS, to serve as criteria to validate initial predictors and as predictors of subsequent Army performance.

2. Improved comprehensive job knowledge tests.

3. Prototype training measures and performance indices to assess components of training success not represented in existing training measures.

4. A procedure for sampling the job content domain to select test content that is more objective and reliable than sampling strategies currently being used operationally.

5. Delineation of the current Army training development and evaluation system.

6. Identification of measures now being administered or recommended for use in training that have the requisite characteristics to be used as surrogate criteria of job performance in on-going development of Army selection and classification procedures and instruments. In addition to development of new measures, Task 3 will yield an evaluation of existing measures as predictors of job performance and as candidate validation criteria. To the extent that general classes of existing

measures are found to be stable across MOS in their predictive validity, the evaluation and classification of existing measures will permit generalizing findings beyond the 19 MOS in Project A.

Scientific Outcomes

1. Determination of the predictive mechanism(s) that explain the relationships between training performance and job performance using two different methods. Predicting job performance using (a) training-relevant only items, and (b) job-and training-relevant items will be used as an indirect means of determining the mechanisms that mediate the relationship between training performance and job performance. Inspection of the unique and common variance in job performance associated with each subset of items will indicate whether learning ability itself (and by inference, learning on the job) or the commonality of elements between school performance measures and job performance measures is the key factor in predicting job performance from training performance.

2. Determination of the adequacy of training performance for validating selection procedures. A fundamental purpose of Project A is to determine whether training performance can serve as an adequate surrogate for job performance in establishing the validity of initial predictors such as ASVAB. Since undergoing training consists of different activities from performing a job, it cannot be assumed that predictors of success in training will also be good predictors of job performance. However, using school performance as the validation criterion achieves

considerable savings in resources compared to validating against job performance. The development of training-relevant and job-relevant training measures in Task 3, therefore, will enable the predictive relationship between school performance and job performance to be determined. From this will come a determination of the adequacy of school performance as an economical validation criterion.

3. Determination of which component(s) of training performance should serve as the focus of predictor development. In addition to establishing overall relationships between school performance and job performance, Task 3 will yield information about various components of school performance--knowledge acquisition, hands-on application, speed of learning, etc.--as they relate to job performance. If speed of learning (a possible motivation measure) is found, in general, to be a better predictor of job performance than knowledge acquisition, for example, this information would guide the development of initial predictors toward that construct.

4. Determination of whether training performance predicts differentially for different groups of trainees (race, gender, mental aptitude) and different groups of MOS (combat, administrative, etc., or other groups). The relationships between training performance and job performance will be analyzed in terms of personal variables and job variables that can be expected to moderate the relationships. For example, school performance may be found to be a better predictor of later performance in jobs that are relatively procedural, such as an administrative specialist, than in jobs that are less structured or

involve more interpersonal contact and decision making, such as military police. Similarly, if training performance is found to be predictive for certain subgroups and not others, then predictors could be selected on the basis of their subgroup correlations, with resulting improvements in accuracy.

REFERENCES

Adkins, D.C. Construction and analysis of achievement tests. Washington, DC: United States Civil Service Commission, 1947.

Anastasi, A. Psychological Testing (4th ed.) New York, NY: MacMillan Publishing Co., Inc., 1976.

Burtch, L.D., Lipscomb, M.S., & Wissman, D.J., Aptitude requirements based on task difficulty: Methodology for evaluation. (AFHRL-TR-81-34) Brooks Air Force Base, Texas: Manpower and Personnel Division, January 1982.

Christal, R.E. What is the value of aptitude tests? Proceedings of the 18th Annual Conference of the Military Testing Association, October 1976.

Cook, T.D., & Campbell, D.T. The design and conduct of quasi-experiments and true experiments in field settings. In M.D. Dunnette (Ed), Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally, 1976.

Ellis, J.A., Wulfeck, W.H., II, & Fredericks, P.S. The instructional quality inventory - II. User's Manual (NPRDC Special Report 79-24). San Diego, CA: Navy Personnel Research and Development Center, August 1979.

Guidelines for development of skill qualification tests. Ft. Eustis, VA: Individual Training and Evaluation Directorate, U.S. Army Training Support Center, December 1977.

Interservice procedures for instructional systems development. TRADOC Pamphlet 350-30, 1975.

Lumsdaine, A.A. Design of training aids and devices. In J.D. Folley, Jr. (Eds), Human Factors Methods for System Design. Pittsburgh: The American Institute for Research for Office of Naval Research, 1960. pp. 217-290. AIR-C90-60-FR-225.

McCain, L.T., & McCleary, R. The statistical analysis of the simple interrupted time-series quasi-experiment. In T.D. Cook and D.T. Campbell (Eds), Quasi-Experimentation: Design and Analysis Issues for Field Settings. Chicago: Rand McNally, 1979.

Mead, D.F. Determining training priorities for job tasks. Proceedings of the 17th Annual Conference, Military Testing Association, 1975, 551-558.

Osborn, W.C. & Ford, P.J., Research on methods of synthetic performance testing. U.S. Army Research Institute for the Behavioral and Social Sciences, April 1976.

Pickering, E.J., & Anderson, A.V. Measurement of job-performance capabilities (NPRDC TR 77-6). San Diego, CA: Navy Personnel Research and Development Center, December 1976.

Smode, A.F., Gruber, A., & Ely, J.H. The measurement of advanced flight vehicle crew proficiency in synthetic ground environments (MRL-TDR-62-2). Wright-Patterson AFB, OH: Behavioral Sciences Laboratory, February 1962.

Vineberg, R., & Joyner, J.N. Prediction of job performance: Review of military studies (NPRDC TR 82-37). San Diego, CA: Navy Personnel Research and Development Center, March 1982.

Vineberg, R., & Taylor, E.N. Performance in four Army jobs by men at different aptitude (AFQT) levels: 3. The relationship of AFQT and job experience to job performance (HumRRO Tech. Rep. 72-22). Alexandria, VA: Human Resources Research Organization, August 1972.

Wernimont, P.F., & Campbell, J.P. Signs, samples and criteria. Journal of Applied Psychology, 1968, 52, 372-376.

Wheaton, G.R., & Fingerman, P.W. Development of a model tank gunnery test (ARI TR-78-A24). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, August 1978.

TASK 4

MEASUREMENT OF ARMY-WIDE PERFORMANCE

GENERAL PURPOSE OF TASK 4

This task is devoted to the identification, refinement, and development of in-service predictors and Army-wide performance measures. In-service predictors are measures obtained after a soldier enters the Army, which predict the soldier's later performance or effectiveness in his/her military career. Army-wide performance measures are those indicators of general performance and effectiveness not related directly to the performance of MOS-specific tasks. This effectiveness domain may also contain measures of a soldier's overall value or worth to the Army.

The central goals of this task are: (a) to identify aspects of soldier effectiveness that apply to all MOS; (b) to identify and/or develop valid indicators to measure these aspects of effectiveness; and (c) to establish the indicators as criteria of soldier effectiveness <u>and</u>, where appropriate, as in-service predictors of future performance or other aspects of soldier effectiveness. Measures must be identified and refined or developed for both first-tour and second-tour performance. In addition, research must determine the utility to the Army of performance levels established by these measures.

Definition of Army-wide effectiveness will require careful specification of the relevant criterion space. "Outcome indicators" and objective administrative indices such as attrition, disciplinary actions, special awards,

schools attended, etc., are clearly Army-wide criteria, and measures of these types of criteria will be of concern in the research. A second focal point will be development of general performance and soldier effectiveness/worth to the Army measures. The "worth to the Army" construct is conceptualized as including a relatively broad set of soldier effectiveness criteria such as organizational commitment, organizational socialization, and morale. Ideally, it is intended to index a soldier's overall value to his/her unit and the Army.

Special behavior-based rating scales will be prepared to measure soldier effectiveness on all important dimensions identified in model development work, and supervisory, peer, and self ratings will be gathered to provide a second set of Army-wide effectiveness criteria. As mentioned, some of these criteria may serve as in-service predictors as well.

BACKGROUND ISSUES AND RATIONALE

Issues in Criterion Development

Obtaining accurate measures of individuals' job performance is absolutely critical in personnel selection research (e.g., Dunnette, 1966; Guion, 1965). Too often, considerable time is spent in developing predictor tests and measures at the expense of: (a) identifying performance constructs that should be targets of predictor measures, and (b) actually measuring in some valid manner the effectiveness of individual performers on those constructs identified. Yet, clearly, test validation results can be meaningful only if proper attention is paid to the "criterion side," so that an accurate depiction of job performance effectiveness is provided. Careful criterion development work should drive identification/development of predictors in selection, and then also provide measures of performance for predictor validation efforts.

Two types of performance measures should be discussed: objective indices, e.g., for an Army clerical MOS - number of pages typed per eight-hour day and number of typing errors made per page, and performance ratings. Objective indices of a worker's performance are in certain cases preferable to the subjective impressions provided by performance ratings, but good objective measures are hard to acquire (Guion, 1965; Landy & Trumbo, 1980).

The difficulty with the vast majority of objective performance measures is that they are almost invariably deficient and/or contaminated (Guion, 1965; Smith, 1976). By deficient, we mean the measure provides only a

partial picture of the worker's effectiveness on the job; that is, there are important aspects of the job left untapped by the objective measure. Referring to the clerical MOS example above, typing speed and accuracy may well be important indices of soldier effectiveness in this MOS, but if helping break in inexperienced typists and willingness to work very hard during heavy production periods are also important for job success, then the former two measures, individually or together, do not adequately measure effectiveness on the job. They are deficient.

Contamination in objective measures occurs when factors that affect how well persons do with respect to the measure are beyond their control. Referring again to the example above, suppose that number of pages typed in a day depends to some extent on the kind of text the typist is to work on, and the soldier has no control over those assignments. The "number of pages" measure, therefore, provides an impure index of effectiveness; it is contaminated. Unfortunately, these are very common problems with objective performance measures. Identifying or developing good, comprehensive objective indices is very difficult.

Our experience with objective indices of soldier effectiveness in the Army, e.g., AWOL, awards, etc., is that individual measures, on their own, are probably deficient as indicators of effectiveness (Borman, Johnson, Motowidlo, & Dunnette, 1975; Shields, Hanser, Williams, & Popelka, 1981). However, composites of these measures formed within a carefully defined conceptual framework may well provide reasonable measures of effectiveness on important Army-wide criteria.

Of course, ratings of effectiveness have their own set of problems. Briefly, factors that lead to inaccuracies in ratings include the following: (a) ratings are often obtained from persons in a poor position to make judgments about incumbent performance (Borman, 1974; Campbell, Dunnette, Lawler, & Weick, 1970); (b) some raters simply lack the observational and/or judgment skills necessary to make accurate evaluations (Borman, 1979); (c) raters often provide biased ratings, based not so much on performance as on race, sex, background, similarity to the ratee in attitudes, etc. (Hamner, Kim, Baird, & Bigoness, 1974; Terborg & Ilgen, 1975); (d) raters sometimes commit rating errors, such as evaluating everyone as very effective, when, in fact, some ratees are performing poorly (Taylor & Wherry, 1951); and (e) raters may fail to use the definitions of the performance dimensions, employing instead their own idiosyncratic beliefs about what it takes to perform effectively, and then rating persons accordingly (Borman and Peterson, in press).

These factors admittedly reflect serious difficulties with ratings, but one feature of rating scale development that should be mentioned is an often overlooked distinct advantage of the method. This feature relates to the ability of a set of well-defined rating dimensions to capture in a comprehensive manner all important performance requirements of a job. That is, if the requirements for successful job performance can be articulated at all, they can be represented in a set of performance dimensions. This means that rating scales, if properly developed, have tremendous <u>potential</u> for generating performance scores that reflect a ratee's actual effectiveness on each important dimension of job performance. It also means that the conceptual definition of a job's performance requirements can be very

well delineated, given the proper approach to rating scale development. In fact, the conceptual definition of job performance, resulting in a set of carefully defined performance dimensions, can provide a framework for criterion measurement in general (both ratings and other criterion indices, such as objective measures).

In other words, the researcher can first gain a definition of the job's performance requirements through rating scale development, and then use both ratings and objective measures to index performance on these important performance requirements. Of course, in this discussion, we are focusing on using rating dimensions as <u>definitions</u> of job performance requirements. A separate question pertains to the ratings themselves on these performance dimensions, and as was mentioned previously, formidable problems are evident with ratings.

On balance, we believe that ratings of soldier performance and effectiveness can be useful in this project. If a few simple but very important principles are followed in gathering performance ratings, the accuracy of these ratings can be maximized. These principles include: (a) developing the rating scales with great care taken to reflect all important performance requirements of the job; (b) creating dimensions that are clearly performance-related and that represent performance factors raters can readily observe in ratees; (c) providing clear, simple directions for using the rating scales; (d) gathering the ratings for research purposes only (rather than for any administrative purpose), and making this clear to the raters; (e) selecting raters who have good opportunity to observe ratee performance, which may mean selecting peers as well as supervisors of the

persons to be rated; and (f) where possible, collecting ratings of each ratee from more than one rater so that interrater agreement may be assessed to provide at least a rough estimate of the accuracy of the ratings. Our experience suggests that when attention is given to such principles, reasonably high quality ratings are likely to emerge (Borman & Peterson, in press).

As mentioned, objective measures may also prove to be useful indices of soldier effectiveness. Relying on objective indices to measure <u>all</u> of the performance criterion domain is unrealistic, we believe. However, targeting selected indices and, especially composites of these indices to measure <u>aspects of</u> performance and effectiveness is likely to be more fruitful. An example might be in tapping soldier effectiveness related to Army discipline. A composite of AWOL, Article 15, and other discipline oriented indicators might serve as a reasonable index of effectiveness in this arena.

In addition, we believe that certain objective, "outcome" indicators might be best used by considering <u>categories</u> of these variables. For example, attrition is a very broad outcome variable in the sense that there are many different reasons for leaving the service. Further, attrition for different reasons, e.g., for medical reasons versus for disciplinary reasons, has very different implications for the kinds of skills, abilities, personal characteristics, etc. that might be relevant as predictors of these outcomes. Accordingly, working with <u>categories</u> of such broad, complex outcome variables (reenlistment is another example) should lead to more conceptually appropriate predictor-criterion links and also to higher

validity coefficients for the predictors. In short, using this category approach to measuring certain outcome variables seems to have more merit for selection research than treating each such variable as a single criterion. Thus, developing/identifying composites targeted toward specific aspects of the criterion performance domain and working with outcome variable categories should make maximally useful the employment of objective indices of performance in this research.

Rationale for Task 4 Research

In our approach, we propose to address the issues and problems of criterion development by building an inductive model of soldier effectiveness, which includes elements of general effectiveness as a soldier and what we have called worth to the Army. By "inductive," we mean that we have no specific hypotheses about the dimensions of soldier effectiveness that may emerge from Task 4 research; we have only preliminary ideas about some of the domains that might be included.

Preliminary hypotheses about the model, elaborated in the Task 4 proposal, suggest that elements represented might include organizational commitment, organizational socialization, and morale. Also to be included in the model are those aspects of soldier job performance that cut across MOS and are therefore important for soldier effectiveness on the job no matter what the specific MOS.

The model is likely to contain multiple dimensions with various personal characteristics/attributes responsible for performance on them. For example, job knowledge might be important for performance or effectiveness

on some dimensions; skills or abilities may be important for other dimensions; and motivation might be important for still other dimensions in the model.

The general idea is for the model's dimensions of soldier effectiveness, derived primarily from behavioral analysis workshops described in the upcoming PROCEDURE section, to provide a framework for development of the actual measures of effectiveness. The dimensions can serve directly as rating scales for superiors and peers to evaluate soldiers in the research. But also, objective measures might be identified or even developed to tap effectiveness on some of the dimensions.

This multimethod approach to measuring performance and effectiveness will be part of a careful construct validation strategy in criterion development work (James, 1973; Smith, 1976). We intend to use the most conceptually appropriate source of performance/effectiveness information to index each element or dimension in the model, but in addition, more than one method, i.e., peer ratings, supervisor ratings, self-ratings, administrative indices, etc., will be employed whenever possible in this measurement effort. Multitrait-multimethod analyses (Kavanaugh, MacKinney, & Wolins, 1971; Lawler, 1967) can then proceed to assess the construct validity of our measures.

Another theme of the criterion development work will be attention to the accuracy of performance and effectiveness measures used in the research project. Recent focus on accuracy in ratings, in addition to psychometric error in these measures (e.g., Bernardin & Pence, 1980; Borman, 1979),

will be attended to during criterion development work. Whenever feasible, ratings and other measures of individual soldier effectiveness will be compared to presumably higher fidelity measures of soldier effectiveness in certain aspects of the job, e.g., Task 5 performance test scores, to enable estimates of accuracy.

SPECIFIC OBJECTIVES

1. Gather/analyze performance and related records data on FY81/82 cohort to aid the Task 1 staff in evaluating the validity of the ASVAB and other available predictors.

2. Develop a model of soldier effectiveness, a conceptual definition of across-MOS soldier performance and worth to the Army.

3. Develop rating scales and scale administration materials for superiors, peers, and self to use in evaluating soldier effectiveness on the model's dimensions.

4. Develop objective composite measures of soldier effectiveness/worth to the Army.

5. Identify attrition and reenlistment categories, e.g., attrition for disciplinary reasons, bars to reenlistment, to serve as outcome criteria of soldier effectiveness/worth to the Army.

6. Develop/identify in-service predictors of second-tour soldier performance.

7. Gather ratings and objective/outcome criterion data on first and second-tour cohorts to aid the Task 1 staff in evaluating the validity of pre-induction predictor measures and the in-service predictors.

8. Obtain scaled utilities for MOS performance levels and jobs.

OVERALL SUMMARY OF THE PROCEDURE

To accomplish these research objectives, the following steps are required:

Information search and evaluation of instruments. We are reviewing the performance measurement literature and examining performance measurement methods currently used in the Army. Also, the Master Plan and Research Plan for this task are being prepared.

Develop prototype instruments. One important activity in this research is the development of a model of soldier effectiveness which will represent a behavioral definition of a broad range of effectiveness dimensions related to across-MOS, Army-wide soldier performance. From the dimensions developed here, we will derive behavioral rating scales to help evaluate individual soldier effectiveness in Task 4 research. We will also develop composites of the administrative indices intended to tap aspects of soldier effectiveness within the model. Finally, we will conduct exploratory research on combat effectiveness on the part of individual soldiers.

FY81/82 cohort records collection, data analyses, and report writing. We will examine records data available for the FY81/82 cohort and collect data on promising records variables for a selected sample of cohort members. Data will be analyzed to evaluate their usefulness as Army-wide criterion measures against which to assess the validity of ASVAB (the latter, a Task 1 activity). Also, we will write draft and final reports on results of these analyses.

Refine existing instruments. This step involves: (a) refining available Army-wide criteria that appear potentially useful for soldier effectiveness measurement; (b) examining attrition and reenlistment categories, reflecting different reasons for leaving the Army and different classes of reenlistment; (c) developing Army-wide measures to tap second-tour soldier (NCO) effectiveness; and (d) developing in-service predictors, first-tour criterion measures that can be used to predict success in second-tour performance.

Prepare for and conduct field tests/revise instruments. We will conduct a series of four field tests on: (a) the first-tour soldier effectiveness rating scales derived from the model, along with the rating scale administrative package; (b) the administrative indices and composites of these indices intended to measure aspects of soldier effectiveness; (c) the second-tour performance/effectiveness measures; and (d) the in-service predictors. For the most part the materials will be revised sequentially, with improvements made in the instruments and supporting materials after each field test period.

FY83/84 cohort first-tour data. Findings in the field tests will result in revised versions of the soldier effectiveness rating scales and a scoring system for the administrative index measures. The scales will be administered to raters evaluating soldiers in this cohort sample, and sample members will also be scored on the administrative indices, including attrition and reenlistment, as appropriate. Criterion data will be analyzed to evaluate their distributions, reliability, convergent and discriminant validity, and, where possible, accuracy in indexing "true"

performance levels. We will also prepare draft and final reports on data analysis results.

Revise instruments. This step will allow for a staff review of the early research on Task 4 criteria, in-depth discussions with ARI and Army officials about the current status of the criterion instruments, and final revisions of these instruments and supporting materials before FY83/84 second-tour and FY86/87 first-tour data collection.

FY83/84 cohort second-tour and FY86/87 cohort first-tour data. We will administer the revised rating scales (first and second tour, as appropriate), score sample members on the administrative indices, and obtain attrition and reenlistment data on appropriate sample members. Data will be analyzed as before, with attention given to distributions of scores, reliability, convergent and discriminant validity, and, where possible, accuracy of these effectiveness measures.

Obtain scaled utilities for MOS performance levels and jobs. The established procedures of multi-attribute utility theory is one approach that can be applied to develop a general model for obtaining and scaling utilities that reflect the relative importance of various MOS at different levels of performance. The model can be implemented in the form of an integrated set of programs run on a microcomputer. These programs present instructions, stimulus material, and assessment and response procedures as appropriate to deriving, analyzing, and maintaining relative utility measures for each MOS and its corresponding performance levels. The software can be used at various data collection centers where appropriate

content area specialists and senior officers are available.

The exact procedures that will be used to obtain the scaled utilities will be worked out through exploratory research. Current plans (see Subtask 4.9) call for accomplishing the scaling in four main steps:

(1) Development of performance construct measures (from those available for each MOS);

(2) Development of utilities <u>within</u> each MOS;

(3) Rescaling the utility of each MOS into a common scale; and

(4) Assigning dollar values to the performance utility levels.

Research will be conducted prior to the completion of each of these steps in order to determine how best to accomplish them.

<u>Prepare final reports</u>. Our staff will prepare draft and final technical reports describing all Task 4 research.

The next section provides a detailed description of the procedural steps to be taken in the Task 4 research. Interrelationships among the research subtasks can be seen in Figure 4-1.

Milestone Chart Task 4

Improve Existing And/Or Develop New In-Service Predictors and Army-Wide Performance Measures

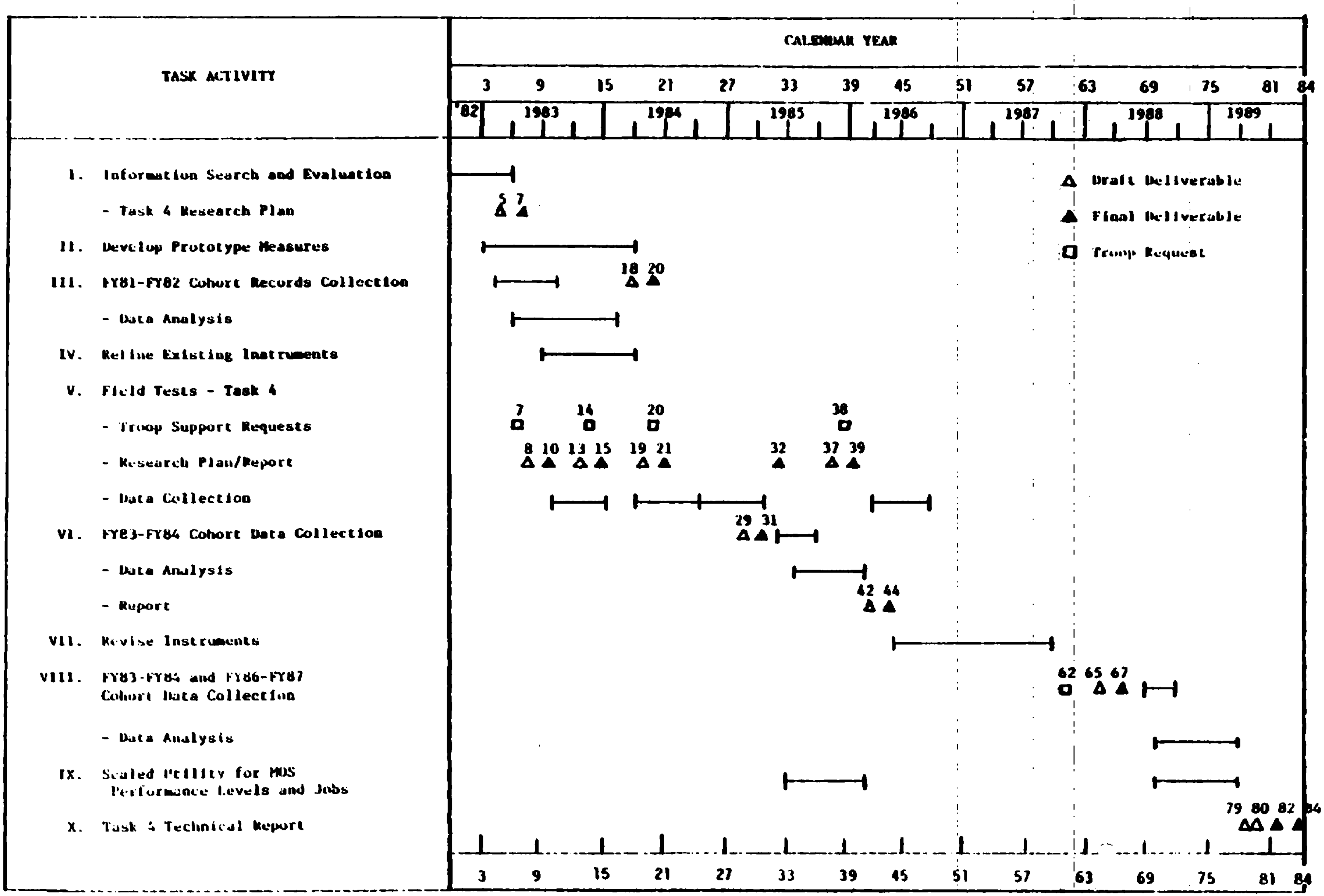


Figure 4-1 Task 4 Milestone Chart

PROCEDURE

Subtask 1: Information Search and Evaluation of Instruments

We are: (a) reviewing the current performance/effectiveness measurement system, (b) conducting a literature review, and (c) developing the research plan and master plan.

Review current system. The first activity in this subtask involves a review of the current performance measurement system in the Army. We are especially concerned with administrative records presently kept on enlisted soldiers. These records may yield performance information to help index individual soldier effectiveness in the Task 4 research.

In order to evaluate their potential usefulness for providing Army-wide criterion data, during the first six months of the research, we reviewed the Enlisted Evaluation Report, the Enlisted Master File (EMF), the Individual Enlisted 201 File, and other record sources available. This activity is in preparation for Subtask 3 interviews with persons familiar with these records sources. Those sources that appear promising in terms of potential for providing meaningful criterion information will be the target of the Subtask 3 interviews.

Conduct literature review. In addition to reviewing current data sources related to potential Task 4 criteria, we are conducting a literature review of books, articles, and technical reports relevant to Task 4. The main purpose of this review is to ensure that no criterion development method,

measuring instrument, or research ideas related to Task 4 are overlooked. We are reviewing both military and civilian research on performance measurement, criterion development and all other topics relevant to Task 4. Performance dimensions surfacing from the military literature will be forwarded to Tasks 2 and 3 as part of a package to be prepared in September 1983.

Compilation and evaluation of cost information. Virtually all validation research is faced with the problem of translating basic information about increments in validity into indices that are the most meaningful and interpretable by the organization. In this project, that will be done by a variety of means. For example, the graphic display of regression functions, decision tables and expectancy charts, and changes in organizational variables such as attrition, reenlistment rates, and frequencies of administration actions, e.g., disciplinary actions, will be used where appropriate.

It should also be possible to attach cost figures to a number of outcomes that are part of the Project A Army-wide criterion assessment efforts. While we do not have the resources to do extensive cost analyses as part of this project, it will be possible to examine information that currently exists or to make use of cost estimation methods already developed by Army managers and other scientists within ARI. Consequently, at about the same time that we begin to develop information on archival criterion information we will conduct interviews with the relevant staff personnel to determine what information is available on the following:

1. Per person training costs for different MOS.
2. Recruiting costs per individual.

3. Costs associated with various categories of attrition.

4. Costs associated with various administrative actions.

If some reasonable set of cost figures can be compiled from the above then it will be possible first to translate validity coefficients into expectancy tables and then to attach cost figures to the reduction in prediction errors that is achieved using a particular set of predictors, or prediction algorithm. These cost figures can then be aggregated over whatever period of time is deemed appropriate. Schmidt, Hunter, and Pearlman (1982) have convincingly argued that in a large organization even small increments in validity can yield enormous cost savings. Depending on how complete they are, the cost figures can also be used in the development of dollar equivalencies for performance utility.

Develop research plan and master plan. Finally, within this subtask, we are preparing the Task 4 Research and Master Plans. The Research Plan describes all proposed project activities, rationale for each activity, troop support requirements for the research, and the scientific and operational outcomes anticipated from the research. The Master Plan details the project staff resources planned for each research activity along with the travel and other direct costs projected for each of these activities.

Subtask 2: Develop Prototype Instruments

This subtask involves two major activities: (a) developing the soldier effectiveness model, and (b) generating composites of administrative indices. Model development steps will be described first and work on the composites will be detailed afterwards.

Development of the soldier effectiveness model. Development of the model will involve primarily an inductive process, consisting of group discussions and workshops with NCO and officers within what has been referred to as the critical incidents or behavioral analysis method (Flanagan, 1954; Smith & Kendall, 1963). In the workshops, we will ask for critical incidents or performance examples describing relatively effective (as well as ineffective) behavioral patterns among first-tour enlisted personnel in a wide range of MOS. The NCO/officer participants will be given special guidance to provide examples that could occur in any MOS, such as the following:

- When on a regular work schedule, this soldier consistently reports for work 15 minutes early and asks the first sergeant if there's anything he can get started on.
- This soldier picked on a fellow unit member by intimidating him in the barracks in front of several other soldiers.

We will also ask members of the workshop groups to tell us about general behavioral patterns that they take into consideration when thinking about a soldier's overall contributions to the Army. We will ask, for instance, how they recognize soldiers whose first-tour performance indicates that they should be encouraged to reenlist for a second tour and other soldiers who should not be encouraged or who should be prevented from reenlisting.

In the behavioral analysis workshops, a brief orientation program (Borman, Hough, & Dunnette, 1976), will be used to train participants in generating and writing useful behavioral examples. In the program, the workshop

leaders will describe the nature of behavioral examples, discuss why generating behavioral examples leads to development of meaningful behavior-based definitions of individual effectiveness, and provide participants with examples of improperly and properly written behavioral examples/critical incidents (e.g., Borman et al, 1976). Workshop participants will then be instructed to begin writing performance examples, and the two research staff members conducting the workshop will help members of the group to ensure they are on the right track. After each participant ha generated four or five examples, we will stop the group and discuss with members the preliminary model of soldier effectiveness that appears in the proposal. This model contains elements of organizational commitment (Steers, 1977), organizational socialization (Van Maanen & Schein, 1979), and morale (Motowidlo & Borman, 1977). The preliminary model is intended to suggest constructs that might be considered as reflecting soldier effectiveness/worth to the Army.

Our staff will seek opinions about components of the model and possible other domains that might be included. However, at no time will we force this model on participants. It will be used only to stimulate discussion about the possible soldier effectiveness domains that might be important to include in the model. After the discussion of domains, participants will be asked to continue writing more examples of effective and ineffective soldier behavior, targeted toward domains they believe to be important for overall soldier effectiveness/worth to the Army. The output from these workshops will be several behavioral examples from each participant.

The examples will be edited into a common format and content-analyzed to form dimensions of Army-wide performance/effectiveness. At least two Task 4 researchers will review the many edited behavioral examples (approximately 1000) and develop categories or dimensions of effectiveness reflecting the content of the examples.

After this step, a "retranslation" process (Smith & Kendall, 1963) provides a method for checking on the dimensions and also a good empirical procedure for developing behavior-based scales to define this performance/effectiveness domain. Specifically, we will present our dimensions to the COR and others he designates to evaluate this depiction of soldier effectiveness. At this point we will also arrange to have the dimensions reviewed by senior field grade officers to ensure that we have adequately captured the domain of soldier effectiveness. We will incorporate suggestions for changes before moving to the retranslation process. Once the revisions are accomplished, we will submit all edited behavioral examples to the workshop participants (by mail), along with the revised dimensions for the retranslation step.

Briefly, with retranslation, participants in this step will sort each example into one of the dimensions according to its perceived content and also rate the effectiveness level of the behavior reflected in the example, e.g., from 1 = very ineffective to 7 = very effective. Retranslation provides a confirmation (or disconfirmation) of the dimension system based on participant agreement on this sorting task. It also provides data on the degree of ambiguity in the behavioral examples' effectiveness levels, so that examples with good agreement in the effectiveness they represent can be used as behavioral anchors for the soldier effectiveness dimensions.

Retranslation, then, results in a series of dimensions, each anchored by scaled behavioral examples. The content of successfully retranslated behavioral examples on each scale is summarized, separately for the effective and ineffective portions of the scale. The final behavior summary scales should reflect, therefore, a clear, behavior-based depiction of the important soldier effectiveness dimensions (see Figure 4-2 for an example behavioral dimension--for the job of Navy recruiter). These dimensions form the content of the model of soldier effectiveness. In addition to representing a very important product for Task 4 research, the dimensions are important for Task 2 predictor development work.

Because Task 2 Subtask 4) requires the dimensions in October, 1983 in order to judge the usefulness of potential predictors, we will forward early versions of them to the Task 2 researchers according to their specifications and in time for their October technical review procedure. The package for Task 2 will include dimensions discovered in the literature review as well as dimensions resulting from early workshops in the model development effort. For each dimension in the package, we will include a name or label, a brief definition of the dimension and any additional explanation related to our view of its meaning, and the source from which the dimension was derived. We will also send this package to Task 3.

Details of troop support for model development. A total of 96 NCO and field grade officers is needed to participate in six one-day workshops (16 at each). We believe NCO and field grade officers will be most appropriate in these and other Task 4 workshops because, typically, they have considerable experience in the Army, while also being reasonably close to the day-to-day operations of Army units and personnel. The NCO/officers should

ESTABLISHING AND MAINTAINING GOOD RELATIONSHIPS IN THE COMMUNITY

Contacting and working effectively with high school counselors, newspaper editors, radio and TV personnel, and others capable of helping recruiters to enlist prospects; building a good reputation for the Navy by developing positive relationships with persons in the community; establishing and maintaining good relationships with parents and family of prospects; presenting a good Navy image in the community.

9 or 10

EXTREMELY EFFECTIVE PERFORMANCE

Is exceptionally adept at cultivating and maintaining excellent relationships with school counselors, teachers, principals, police, news media persons, local business persons, and other persons who are important for getting referrals and free advertising.

Is innovative in informing the public about the Navy; actively promotes the Navy and makes friends for the Navy while doing it; always distributes the most current Navy information.

Volunteers off-duty time to work on community projects, celebrations, parades, etc.

6, 7, or 8

EFFECTIVE PERFORMANCE

Spends productive time with individuals such as police, city government, or school officials; may lunch with them, distribute calendars, appointment books, buttons, etc., to them, and/or invite them for cocktails.

Arranges for interested persons such Navy activities as trips to the Naval Academy; keeps relevant persons informed of Navy activities.

Encourages principals, counselors, and other persons important to a prospect to call if they have any questions about the Navy.

3, 4, or 5

MARGINAL PERFORMANCE

Contacts school officials only sporadically; keeps them waiting for information they want; relationships with counselors, teachers, etc., and persons important to an applicant or recruit are distant and under-developed.

Is not alert to opportunities to promote the Navy; rarely volunteers off-duty time to promote the Navy and is unenthusiastic when approached to do something for the community; rarely accepts speaking invitations.

Is, at times, discourteous to persons in the community; for example, sends form letters to persons who have assisted him or other Navy recruiters; is not always alert to the family's desire for more information about the Navy and the program in which their son or daughter enlisted.

---

1 or 2

INEFFECTIVE PERFORMANCE

Does not contact high school counselors; does not accept speaking engagements; drives around in car instead of getting out and meeting people.

Alienates persons in community or persons important to an applicant or recruit by ignoring them, not answering their questions, responding rudely, demanding information, encouraging high school students to drop out of school; sometimes does not appear at recruiting presentations for which he/she is scheduled.

Presents negative image of the Navy by doing things like driving while intoxicated or speeding and honking impatiently at other drivers; may express dislike for the Navy or recruiting.

Figure 4-2

Behavior Summary Scale for Job of Navy Recruiter

be selected to ensure that they are interested and willing to participate. The mix of MOS and officer specialties reflected in the groups is not particularly important, as long as a wide variety of MOS/specialties is represented across the six participant groups. This configuration should prevent the model from being a narrow, parochial definition of Army-wide soldier effectiveness.

The number of participants was carefully considered and should lead to generation of the proper number of behavioral examples. Based on our experience with behavioral scale development, 800-1000 examples are required to sample the performance domain sufficiently to develop rich behavioral definitions of the domain (Borman, Hough, & Dunnette, 1976). Also our experience shows that each of the 96 participants, working with us for one day, should be able to generate about 10 usable examples (Borman, Johnson, Motowidlo, & Dunnette, 1975).

The total time required of each NCO/officer will be one and one-half days. He/she must attend the one-day workshop to generate behavioral examples and later must respond to the retranslation task, which requires reviewing our dimensions identified for the model and making ratings of effectiveness levels that were suggested by each behavioral example written in the workshops. The review and ratings should take one-half day of his/her time.

Recommended locations for the workshops are Forts Hood, Knox, Benning, and Carson. This site selection should result in a representative mix of MOS/officer specialties with relatively little travel on the part of participants.

Proposed timing of the workshops is as follows:

- o Last two weeks in July, 1983: Workshops 1 and 2 (Hood)
- o First two weeks in Sept, 1983: Workshops 3 and 4 (Knox & Benning)
- o Last two weeks in Sept, 1983: Workshops 5 and 6 (Carson)

It is important to have one month between the first two workshops and the later workshops to enable the Task 4 staff to: (a) examine the initial behavioral examples to ensure that they meet project requirements, and (b) form preliminary dimensions based on the content of these examples. As mentioned, these dimensions and definitions will be forwarded to Tasks 2 and 3.

The schedule presented should allow sufficient time for this dimension development effort. At the same time, the schedule is "tight" to ensure efficient use of staff and relatively quick completion of the model development steps.

Another activity related to model development is preparation of (a) rating scales based on the model, and (b) a rating scale administration package to aid in gathering rating data. Developing the rating format is very straightforward. The dimensions, including behavioral definitions, emerge directly from the model development steps, and therefore no additional work is required to ready the rating scales.

For the administration package, we envision developing instructions that enable raters to complete their ratings with maximum ease and minimum confusion. We will use our past experience in studies involving ratings

(e.g., Borman & Dunnette, 1975; Campbell, Dunnette, Arvey, & Hellervik, 1973) to prepare the best instructions possible, and then research will be conducted during field tests (described later in this plan) of the rating form and procedures.

Also to be developed is a rater training module to help raters make more accurate evaluations. As with the rating scale instructions, we will develop what we believe to be the best training module possible using our experience in past research (e.g., Borman, 1975; Peterson, Houston, & Rosse, 1981). Briefly, the module we have worked with describes three different rating errors - halo, stereotype, leniency/restriction-of-range - and instructs raters on the use of behavioral definitions when making their evaluations. This is presented on one or two pages and is easy for the lay person to understand. The trainer also discusses these errors in common sense terms, assures the raters that the evaluations will be for research-only purposes, and answers questions about the rating form and the project in general. As with the administration package, we will conduct field tests of this and similar modules to improve further the rater training component.

<u>Pilot development of special Combat Performance Prediction Scales</u>. The major rating scale development work in this subtask will focus on the model development steps just detailed. However, we plan an exploratory investigation to determine the feasibility of constructing Combat Performance Prediction Scales that might be used to predict performance in combat. Such scales would be designed to evaluate expected performance under the degraded environmental conditions, communication, support, etc., and the

increased confusion, workload, and uncertainty of a combat environment. Such conditions would be expected for many soldiers near a battle area, even though it is likely that only a small percentage of the total Army force will directly participate in combat. Thus, soldier effectiveness under combat conditions represents a potentially important part of the total soldier effectiveness domain.

Still another reason for considering this an exploratory effort and for concentrating on the soldier effectiveness domains described previously, i.e., across-MOS performance, organizational commitment and socialization, etc., is that we may be asking raters to perform a rating task they are incapable of doing well. With these scales, we will be asking raters to evaluate the likely performance of individual soldiers on dimensions relevant to a combat situation. This requires considerable inference on the part of the raters, because we are asking that they observe garrison/field performance and effectiveness and provide estimates of effectiveness in a very different setting.

Nonetheless, we plan to conduct this exploratory work in an attempt to form dimensions of combat performance and to develop the Combat Performance Prediction Scales. These scales, like the soldier effectiveness model dimensions, will be appropriate for any MOS. The dimensions will be developed in three one-day workshops with a total of 30 senior NCOs and field grade officers participating in each session (10 in each). These participants should have experience in combat environments.

In each workshop, two Task 4 researchers will describe the objectives of the session and then ask participants to consider effective and ineffective soldier performance in combat. Participants will be encouraged to describe to the group what kinds of performances seem to differentiate those who prove to be effective in combat and those who are less effective. During the group discussion we will guide participants toward identifying and defining dimensions of combat performance. In the second and third workshops, we will also present to the groups what we have gleaned from earlier sessions. Participants will be asked to comment on the dimension names and definitions and to make revisions as needed. The product to emerge from this series of sessions will be a set of well-defined dimensions that can be used as "predicted performance" rating scales to predict the combat performance of individual soldiers.

The three workshops should be conducted at the Pentagon. As before, workshop participants should be selected to ensure they are interested in the project and well motivated to participate. We strongly suggest 10 participants in each workshop because this is a small enough number to get everyone's point-of-view and yet large enough to allow (across the three workshops) reasonably good representation in terms of specialties.

Development of administrative index composites. Regarding this development work, Subtask 1 activities should yield candidate indices for these composites. Interviews conducted in Subtask 3 will suggest which indices appear most promising for further examination, and the preliminary analyses of records data, also accomplished in Subtask 3, should provide more definitive guidance on the usefulness of individual administrative indices.

However, as argued in our proposal, we see considerable merit in forming composites of selected indices to tap elements of the soldier effectiveness model, and work in Subtask 2 is intended to accomplish this objective.

Briefly, the argument is this. There is a serious difficulty in using administrative records as soldier effectiveness criteria since they often reflect exceptionally good or exceptionally poor performance. These records, therefore, have low base rates, i.e., they appear infrequently in a soldier's records, and very little variance, i.e., everyone has about the same "scores" on them. Consider, for example, AWOL on the poor performance side and special awards on the good performance side. This skewness seriously constrains the usefulness of an administrative variable as a criterion of soldier effectiveness (cf. Hammer & Landau, 1981).

One strategy for dealing with the problem of low base rates is to combine records of different kinds of events and actions into more general indices. An approach to doing this is to examine patterns of correlations between different records--provided, of course, that there is enough variance to permit at least some patterns of covariation to emerge--and combine those that are empirically related in this fashion. A second way, which might still be possible even though empirical relationships between records cannot be detected, is to combine different elements that are judged to be conceptually similar. Quite possibly, for example, it may become clear that several kinds of awards should be combined into one index because they all indicate organizational recognition for outstanding performance in some psychologically homogeneous behavioral domain. Then a soldier's "score" would be the total number of awards received in that particular category.

How this combining of individual objective indices might offer a good approach can be explored using data recently presented by Shields, Hanser, Williams, and Popelka (1981). They gathered information on soldier effectiveness in the 193rd Infantry Brigade, Panama. Data were collected on such variables as SQT scores, number of awards, number of military courses completed, number of times honor graduate status was attained in training courses, number of Article 15, and number of letters of appreciation. One result of the research was that positive correlations emerged between some criterion pairs--for example, SQT scores and number of awards, r=.43; number of awards and number of military courses completed, r=.63; etc. This indicates that these different indices may indeed reflect to some extent an underlying effectiveness construct. Of course, relationships between other pairs of indices are low, but what we suggest here is that low base rates may be an important contributor to the low correlations in some cases. For example, "number of times honor graduate status attained" has a mean of .03 across some 125 soldiers, and this low base rate reduces the likelihood of substantial correlations between this variable and other variables. However, if scores on this measure are combined with scores on other low base rate but conceptually similar measures, i.e., measuring what appears to be the same underlying construct, the base rate might well improve to a level where significant correlations with other variables would be much more likely.

A final issue here is the problem of weighting different records so that the more important or organizationally significant pieces of information will play a heavier role in determining a soldier's total score on an aggregated index. We must, essentially, seek to determine how many

"points" a soldier should get for each of several kinds of awards, for example, and how many "points" should be taken away for AWOL, Article 15, and different punishments. We hope, then, to develop a rationale for constructing one or more weighted indices of soldier effectiveness in the context of Army-wide performance. Indices generated in this manner will probably show more variance than the raw individual records that make them up, and they should, therefore, be more useful as criterion measures of Army-wide performance.

To accomplish this step, we will prepare for and then conduct two different one-day workshops with officers and NCOs (details of troop support requirements follow). In preparation for these sessions we will develop a briefing package that describes in lay terms the statistical strengths and weaknesses of each candidate effectiveness index. Base rate and missing data information, correlations between indices, when available, and any other kind of information from the Subtask 3 analyses of records data will be put in a form that will give workshop participants a good picture of each variable's usefulness and how the variables relate to each other empirically.

Specifically, two research staff members will introduce the mission for the workshop series to form one or more composites reflecting important constructs of soldier effectiveness. Then the staff will explain the package describing statistical properties of the candidate indices and begin a discussion of what kind(s) of composites might be formed to tap important aspects of the soldier effectiveness domain.

The main idea for these workshops is that the participants will have considered opinions about the meaning of different objective indices and the importance of each (conceptually) as an index of some aspect of soldier effectiveness, and we as psychologists will have good knowledge of the measurement properties of these candidate indices. These appear to be exactly the two kinds of information and knowledge necessary to make good judgments about forming the composite or composites.

Thus, we envision the officer and NCO participants offering their views on the measures to go into the composite(s) and the Task 4 staff (after seeking counsel from Task 1 researchers) providing statistical guidance to ensure that the composite(s) is formed on a reasonably sound psychometric basis. Workshop participants will be asked also to provide weights for the individual elements of the composite(s), again, based on a combination of conceptual and statistical considerations. The final composites and weights will be formed according to a consensus of the final opinions expressed in each of the three workshops.

In sum, the output from the workshop series will be: (a) labels and conceptual definitions of one or more composites targeted toward measuring one or more important Army-wide criterion constructs; and (b) member objective indices for each composite, along with weights assigned. All information regarding these indices will be forwarded to Task 2 scientists.

<u>Details of troop support for development of administrative index composites</u>. A total of 20 NCO and field grade officers (10 of each) will be assigned to participate in two one-day workshops. As with the behavioral

analysis workshops, participants should be selected to include well motivated and knowledgeable officers and NCOs from a variety of specialties and types of units. This is to preclude obtaining a narrow view of these administrative measures. We recommend the number 10 in each workshop because this is a small enough group to ensure everyone is heard from, and yet large enough to allow (across the two workshops) reasonably broad representation in terms of specialties.

The workshops should be conducted at the U. S. Army in Europe in October, 1983. At that point we will have learned enough from Subtask 3 records research to prepare the briefing materials referred to previously.

Subtask 3: FY81/82 Cohort Records Collection, Data Analyses, and Report Writing

We will: (a) examine records data, (b) collect records data on the FY81/82 cohort, (c) perform data analyses on FY81/82 cohort records, and (d) prepare draft and final reports on FY81/82 cohort analyses.

Examination and collection of records data. The first two activities in this subtask involve the examination and collection of records information on the FY81/82 cohort. Accordingly, we propose to examine and collect records data from the Enlisted Master File (EMF), the Individual Enlisted 201 File, and any other record sources available as identified in

Subtask 1 to evaluate their usefulness for providing Army-wide criterion data. To accomplish this, we will conduct 20 two-hour structured interviews with persons familiar with one or more records source. The focus of these interviews will be on the state of the data (How much missing data, extent and types of error in data, etc., exist?) and on the meaning of the information (How precisely is each attrition category defined?). Task 4 researchers have already conducted several of these interviews at FORSCOM headquarters, Ft. McPherson, and with MILPERCEN officials in their Alexandria, Virginia location.

Data analyses. Another very important step in assessing the state of available records is to perform preliminary data analyses on the FY81/82 cohort to determine: (a) amount of missing data; (b) base rates of the variables we are concerned with in the research; (c) psychometric characteristics of the measures, e.g., means, standard deviations, across-time reliability, etc.; and (d) possible serious discrepancies between MOS in base rates, means, and standard deviations of measures (as appropriate). The general idea is to check closely on the "quality" of the data. High rates of missing data, very uneven base rates, poor across-time reliability where we would expect consistency, and large differences between MOS in average "scores" on criterion measures will be cause for concern. However, if one measure proves to be a problem in this regard it is possible that another measure in the same domain or a similar domain might be substituted. Whatever the exact outcome here, the intent is the same as with the interviews conducted in Subtask 1 - become as knowledgeable as possible about the data available on Army-wide criteria so that subsequent analysis results are properly interpreted.

Specifically, as stated in the proposal, we plan to perform the above initial and exploratory data analyses on the following measures: attrition, including different categories of the variable, e.g., medical, drug-related, etc.; reenlistment, including bars and reenlistment choice; Article 15 and courts-martial; promotions; school selection; and AWOL. Visits to Ft. McPherson (FORSCOM) and MILPERCEN have revealed that information on attrition, reenlistment eligibility, and bars to reenlistment is available from the EMF. Information on awards, Article 15, letters of commendation, etc., is not, however, available from the EMF. This information, which exists on microfiche for all enlisted personnel, is centrally located at Ft. Benjamin Harrison, Indiana. Additionally, data analyses will be performed on any other index available in the records, as identified in the structured interviews, and which might be an indicator of important components of soldier effectiveness.

Thus, we will visit Ft. Harrison to review 400 microfiche records on individual soldiers to help evaluate the usefulness of these records. Analyses will be performed to evaluate base rates, the amount of missing data, etc. What is learned from these analyses can be applied both to research with the FY81/82 cohort <u>and</u> to subsequent work on objective measure composites.

After the records data have been examined and decisions made concerning the most promising variables, we will actually score each member of the FY81/82 cohort sample (all those cohort members selected for the validation research) on each of these variables. These scores will then serve as criterion data in the ASVAB validation work to be conducted by Task 1 scientists.

After the criterion records data have been added to the LRDB by Task 1 staff, we will perform appropriate analyses on the data. These are likely to include estimates of reliability where possible, correlations between variables, and factor analysis work to evaluate the criterion dimensionality.

Prepare draft and final report on FY81/82 cohort data analyses. Finally, within this subtask, draft and final reports on the FY81/82 cohort analyses will be prepared. These reports will summarize the findings from data analyses, and as needed, make recommendations for improving the collection and recording of objective indices of soldier effectiveness.

Subtask 4: Refine Existing Instruments

We will: (a) suggest revisions in Army-wide criterion measures identified in Subtasks 1, 2 and 3 in an effort to improve them, (b) examine categories of attrition and reenlistment criteria to improve predictor-criterion match-ups, (c) develop second-tour performance measures, and (d) develop in-service predictors.

Revising existing Army-wide criteria. First, our staff will analyze carefully the problems with Army-wide criteria discovered in Subtasks 1, 2 and 3 of Task 4, and, as appropriate, suggest improvements in the measures. Refinement of Army-wide measures may take a number of different forms depending upon identified requirements and the instruments or procedures concerned. Refinements might involve nothing more than changes in records

forms or reporting requirements for personnel actions to improve the quality of obtained data. On the other hand, refinement could require the revision of forms, instruments, or procedures, with significant impact ultimately upon current administrative procedures. Throughout, potential impact upon administrative procedures and requirements will be an important consideration. The objective will be to avoid, if possible, or to minimize: (a) potential changes in administrative or reporting requirements that might cause problems for operating units, and (b) additional effort by personnel of operating units. The main point of this activity, however, is to refine forms, instruments, and procedures so that administrative index data are of the highest quality possible.

We propose to interview 20 persons knowledgeable about administrative records during the Subtask 3 research. Within Subtask 4, it will be important to return to 5 of these same individuals to review with them our suggested revisions to records forms, instruments, and procedures for collecting administrative data (we recommend January - February 1984). These reviews are to evaluate the practicality of our suggested revisions. There is no intent by the contract researchers to institute these changes. The above mentioned activity is meant to provide information that might generate recommendations for more efficient and productive collection of administrative records data.

<u>Examine categories of attrition and reenlistment</u>. A second effort in this subtask involves studying categories of attrition and reenlistment to develop more homogeneous criteria. Earlier, we discussed the notion of creating more conceptually sensible predictor-criterion linkages by reducing the global attrition criterion to a series of criteria, each one

homogeneous in terms of the likely reasons for the outcome, e.g., discipline problems, medical reasons, etc. With this approach, correlations between pre-induction predictors and categories of attrition are likely to be much higher, and make better conceptual sense than if attrition as a whole is used as a criterion. Having spoken with personnel at Ft. McPherson (FORSCOM) and MILPERCEN, we have learned that attrition, reenlistment eligibility, and bars to reenlistment are currently recorded by category on the EMF.

Thus, we plan to study the attrition and reenlistment categories to determine the frequency with which they are being used. An example of such a category is attrition for discipline-related reasons. It is likely that pre-induction predictors measuring socialization and adjustment factors may be successful in predicting this kind of attrition. Likewise, other such categories will be examined using this concept and these categories operationally defined for tryout in the field tests. The same approach will be used to evaluate the reenlistment categories. If, however, we find that particular categories are not being used, we will form composites of categories in the same manner that composites of other administrative indices are formed. For example, attrition due to AWOL and Article 15 may be combined to form a conceptually homogeneous composite called Disciplinary Attrition.

Developing second-tour performance measures. A third activity in this subtask is developing second-tour performance measures to serve as criteria both for pre-induction predictors and for in-service predictors. One important aspect of second-tour performance is the leader behavior of

second-tour soldiers serving as NCO. Leader behavior may involve such activities as supervising, training, counseling, planning, decision making, problem solving, etc., all of which are common to and required of all NCO, regardless of MOS.

The proper approach to measuring such leader behavior will be the identification of dimensions of leader performance common to all NCO, and then the development of measures of the dimensions. Fortunately, it will not be necessary to go through the process of empirically deriving such dimensions in this project. As part of a recently completed project in USAREUR for ARI, HumRRO personnel have already identified the dimensions of leader behavior for four levels of NCO, one of which is squad leader, the first NCO level (Hebein, Kaplan, Olmstead, & Sharon, 1983). Therefore, as a starting point, we plan to evaluate the usefulness of the dimensions derived from the USAREUR project as one set of variables for measuring second-tour performance. Additional dimensions may also be developed as part of this project.

We will examine carefully the dimensions of soldier effectiveness derived for the model prepared in Subtask 2 to evaluate their appropriateness for inclusion as second-tour performance dimensions. The dimensions from the ARI-HumRRO work along with selected dimensions from the model will then be presented to three workshop groups of 15 NCO and field grade officers in each workshop. The dimension set will be refined on the basis of their suggestions. At these one-day workshops, to be held during January-February 1984 at Forts Hood, Bragg, and Benning, participants will also be asked to provide behavioral statements to anchor the effective and

ineffective portions of each dimension. Subsequent to the workshop, our staff will refine the participants' input and submit the scales, including the behavioral anchors, to the COR and other ARI scientists, as well as to designated Army personnel, for review and suggestions. We will revise the scales based on these reviews. Thus, behavior-based rating scales will be developed for second-tour NCO performance, based in part on work done in USAREUR.

Again, the number of participants for these workshops was carefully considered and seems appropriate. We will already have behavioral dimensions from the first-tour behavioral scale development effort, as well as input from the USAREUR study. Therefore, 45 participants (one-half the number being used in the earlier scale development work) should be sufficient to generate the needed behavioral examples.

Developing in-service predictors of second-tour performance. This effort will rely heavily on an underlying model of behavioral consistency (Wernimont & Campbell, 1968)--that is, the best predictor of future behavior in a domain is past behavior in that same domain. Accordingly, first-tour in-service predictors will be identified or developed based on a conceptual match between first and second-tour criteria. For example, in the area of discipline-related second-tour criteria, we will seek first-tour predictors that index the same kinds of behaviors.

Specifically, we will form hypotheses about in-service predictors/second-tour performance links, drawing on the rating scales developed in the model of soldier effectiveness and the composite(s) of objective indices as the

main in-service predictors. These hypotheses will be formally stated so that we can later check on the validity of the in-service measures for predicting second-tour soldier effectiveness (Task 1 activity). Again, behavior consistency notions will drive development of these hypotheses. Evidence of first-tour leadership (as possibly indexed by ratings on certain dimensions of the model) will be used as a predictor of second-tour leadership effectiveness. Getting in trouble during the first tour should predict discipline-related criteria in the second tour, and so on.

Subtask 5: Prepare for and Conduct Field Tests and Revise Instruments

The purposes of this subtask are to: (a) try out, under operational conditions, the Army-wide criterion measures developed to date; and (b) revise these measures as a function of shortcomings which arise. It is our intent to refine Task 4 measures through an iterative process of four field test cycles. These revisions are expected to include changes in both content and format. The fourth field test will be conducted in FY1986, after the FY83/84 cohort first-tour data have been collected and analyzed (Subtask 6).

Each field test cycle consists of three activities. The first of these is the development of a detailed field test plan. The plan will contain a rationale for collecting the data, copies of the data collection instruments themselves, and proposed data analyses. The second and third activities are the field test itself followed by analysis of the results.

Field test data collection. The design of any one field test can be expected to vary as a function of prior field tests. ~wever, all field tests will be related to the solution of problems which follow from

attempting to measure the performance of a large number of individuals as accurately as possible. Several questions, some more critical than others, arise with regard to this effort. It is our intention to address the more critical issues in the early field tests so that any major problems which develop can be dealt with in a timely manner.

There are three issues which we believe are most important for the success of this subtask, and as a result will be examined in the first field test: (a) the psychometric quality and accuracy of the resulting measurements; (b) the applicability of the measures across disparate Army occupations; and (c) the acceptability of the instruments to the Army. The first field test is designed to evaluate the extent to which our instruments meet these criteria. Another issue to be explored in a later field test will be an examination of the differences in rating distributions that may be attributable to dropping the "for research purposes only" phrase from the rating instructions.

Field tests are designed around "rating units." Each rating unit consists of the individual soldier to be evaluated, two identifiable peers, and two identifiable supervisors. A peer is defined as an individual assigned to the same platoon as the individual. A supervisor is defined as the individual's platoon sergeant or platoon leader.

In order to collect sufficient appropriate data to address the critical issues raised above, the following data will be collected on each rating unit: (a) minimum of two peer evaluations using prototype scales, (b) minimum of two supervisory evaluations using prototype scales, (c) length

of time known by each rater, (d) self-ratings using prototype instruments, and (e) objective/administrative indices developed in Subtask 4.2.

MOS to be used are the ones that will be tested initially by Task 5 personnel (13B - Cannon Crewman, 71L - Administrative Specialist, 95B - Military Police, 64C - Motor Transport Operator). Thus, in addition to the data collected under the auspices of Task 4 staff, other criterion data should be available on the same individuals from Task 5 research. This is extremely important, for it allows a determination of the convergent and discriminant validity of criterion measures across these two tasks. Obviously, it will also result in a reduction of data collection costs and disruption of ongoing troop activities.

In addition to the above, we also plan to collect evaluative information from the supervisory raters concerning their reactions to the rating scales themselves. This evaluative component will be augmented by brief oral interviews with approximately one-fourth of the supervisory raters, again focusing on their reactions to the scales.

Because of statistical power considerations (Schmidt, Hunter, & Urry, 1976) and likely missing or incomplete data, an initial sample of 150 rating units in each of the four target occupations is projected. If we assume a span of control of 5-10 subordinates per supervisor and the required two supervisors per individual soldier, a total of 40 supervisors and 150 soldiers representing each MOS will be required for the first field test. Each of these individuals will be needed for a one day period. (During this time the Task 5 job-specific rating scales will also be administered.)

Access will also be needed to each of the 150 soldier's 201 Personnel File. While not yet projected, sample sizes required for each of the remaining field tests are expected to be comparable.

Analyzing the field test data. These analyses will include:

1. Examining the distributions of ratings and administrative index score - We will evaluate the distribution, i.e., means, standard deviations, skewness, of the ratings and also the administrative index scores. With the ratings, for example, we will have certain expectations about how the distributions should and should not look, and the actual distributions will be compared to our preconceptions. Severe deviations from such expected distributions would cause concern.

2. Evaluating the interrater agreement of ratings - Although it is possible to obtain high interrater agreement and still have very inaccurate ratings, good agreement between raters providing independent performance judgments is generally thought of as a positive sign concerning ratings. We will compute the interrater agreement both within rating source, i.e., between supervisors and between peers, and across the two sources to help assess the quality of rating data.

3. Examining the dimensionality of ratings and the administrative data - Employing factor analysis methods, we will evaluate the dimensionality, i.e., factor structure, of the rating data and probably the rating and administrative index data together. It may be, for example, that a technical competence versus interpersonal competence/adjustment to the Army structure will emerge from factor analytic work. Obtaining two or

more reliable and psychologically meaningful factors from criterion data would be very encouraging for the validity analyses. This is because very different kinds of predictor measures are likely to be successful in predicting soldier effectiveness in very different parts of the effectiveness domain, and the emergence of such reliable factors makes it possible to study these relatively refined predictor - criterion linkages. As an example, we found that a mechanical comprehension test correlated higher with ratings of technical competence (on transmission and distribution jobs) than it did with ratings of interpersonal effectiveness on the job (Borman, Mendel, Lammlein, & Rosse, 1981).

4. Evaluating the convergent and discriminant validity of the ratings and the ratings and administrative indices together - We favor an analysis strategy suggested by Kavanagh, MacKinney, and Wolins (1971) to evaluate rating data. This strategy yields estimates of convergence across rating source (essentially interrater agreement) <u>and</u> the discriminant validity of ratings (how reliably raters evaluate different aspects of soldier effectiveness). This method provides good information on the quality of ratings that can be reasonably compared across settings, e.g., units. It may be possible, also, to evaluate convergence (or across-method reliability) of the administrative index data and the rating data, using this data analysis strategy.

5. Assessing accuracy of the ratings and administrative index scores - The Task 5 performance measures will presumably offer high fidelity indices of soldier effectiveness in some parts of the job. Therefore, for the

nine MOS for which both sets of data have been collected it may be reasonable to correlate Task 4 ratings and administrative index scores against selected Task 5 performance scores to assess the accuracy of the Task 4 measures.

We must be careful here because we would not expect high correlations between Task 5 technical performance scores and rated effectiveness in the area of organizational commitment, for example. On the other hand, certain technical competence dimensions derived in Task 4 would be expected to correlate with performance scores generated in Task 5. Thus, where the intended content of Task 4 and Task 5 measures are the same or very similar, correlations between the two sets of scores may provide a meaningful estimate of the accuracy of Task 4 measures.

6. Finally, we will consider applying generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) to analyze the ratings. Using this approach, we can identify sources of variance that may affect the distributions of ratings and categorize these sources into desirable or true sources of variance and undesirable or error sources. An attractive feature of this approach is that it specifies analyses of variance that could provide information on two important potential sources of error in ratings - MOS and location/unit effects. If these analyses show that substantial variance in ratings is due to these effects, it suggests that raters in different MOS and units are using the rating scales very differently, and therefore, analyses cutting across MOS/units are suspect. Care taken in developing and administering the

scales will serve to minimize these effects, but the generalizability theory approach provides a possible method for checking on this potential problem.

Still another very important evaluation criterion for the field test is "user acceptance," which relates to how smoothly the data collection proceeds. Recall that we plan to administer questionnaires to raters and interview them after they complete their ratings. We will evaluate the questionnaire and interview responses, with an eye toward revising our procedures if it seems warranted. Also, we will, of course, carefully monitor the data-gathering procedures, both ratings and objective measures, to evaluate the feasibility of these procedures for the larger-scale data collection efforts.

<u>Pilot test of special Combat Performance Predictor Scales</u>. As mentioned, we plan to explore the feasibility of developing performance prediction dimensions and rating scales relevant to a combat/mobilization context. These dimensions will reflect across-MOS performance requirements in a combat situation. In a pilot field test of these rating scales, we will ask 50 rating units (20 supervisors and 20 peers of the 50 target ratees) in each of the four MOS to be included in the initial field test to make ratings of soldier effectiveness using both the combat performance prediction scales <u>and</u> the rating scales developed to evaluate non-combat Army-wide soldier effectiveness. The primary goal will be to assess raters' abilities to differentiate between non-combat performance and predicted combat performance. One index of this will be a comparison between inter-rater reliability for the prediction scale evaluations with the level of

reliability attained on the non-combat rating scales. Additionally, we will interview one-fourth of these raters (20 supervisors and 20 peers) to obtain their impressions of the predicted performance rating task. They will be asked about how they made their prediction evaluations, how confident they felt about the predictions, and how the rating scales and/or rating procedure could be improved.

In sum, our staff will evaluate the data collection procedures and perform data analyses in an effort to identify weaknesses in the measurement system. The total evaluation effort will suggest which instruments and procedures must be improved upon; necessary revisions will be made, and revised measures will be field tested.

Prepare report on field tests I, II and III. Finally, within this subtask, a report on the first three field tests will be prepared. The report will summarize the findings from data analyses and describe the revisions made to measures and procedures as a result of each of the three field tests. Additionally, the report will outline the revisions, if needed, to procedures and measures that will be evaluated in the fourth and final field test in FY 1986.

Subtask 6: FY83/84 Cohort First-Tour Data Collection and Analyses

This subtask consists of four major activities: (a) preparation of a draft and final plan for the FY83/84 cohort data collection, (b) administration of the rating and administrative measures to the sample of cohort members in the 19 MOS, (c) analysis of data from this data collection, and (d) preparation of a draft and final report on these analyses.

Draft and final plan for data collection. The Task 4 staff will prepare a draft plan for data collection from the FY83/84 cohort and submit the plan to the COR for review. The plan will discuss rating scale administration procedures and collection of objective administrative index data. After review by the COR and others he designates, we will revise the draft plan and submit a final version.

Administration of the rating and administrative measures. The HumRRO-PDRI staff members will monitor collection of the rating and administrative data. Although Task 5 researchers will for the most part be performing the actual data gathering work, Task 4 researchers will be on hand initially to ensure that the administration procedures are being handled properly. We are especially concerned about rating scale administration. It is critical that careful attention be directed to several aspects of this administration effort. Selecting the proper supervisor and peer raters (according to their knowledge of ratee performance), introducing the study in a professional but motivating manner to encourage complete cooperation in the rating sessions, and training raters to increase the likelihood of obtaining accurate evaluations are all very important. These and other details of the scale administration procedures will be closely monitored by our staff.

Subject to modification based on field test results (Subtask 5), we will ask two supervisors (first sergeant and platoon leader) and two peers of the target ratees to complete the performance rating forms on soldiers they are to evaluate. Rating assignments will be determined before the actual scale administration sessions in discussions between researchers and the

first sergeants of the ratees. The rating sessions will be conducted and the data processed and put into the LRDB by Task 1 personnel along with the administrative indices data.

For attrition and reenlistment categories we will obtain these data on those soldiers in the FY83/84 cohort sample who have either separated or reenlisted and, again, get the information processed and entered on the LRDB.

Data analyses. Task 1 staff is responsible for validity analyses, but Task 4 researchers will perform analyses to determine the quality of the data and the instruments. These analyses will include the same ones discussed for Subtask 5: (a) evaluating the distributions of ratings and administrative index scores; (b) evaluating the interrater agreement of ratings; (c) determining the dimensionality of ratings; (d) evaluating the convergent and discriminant validity of the ratings; (e) assessing, as feasible and appropriate, the accuracy of the ratings; and (f) evaluating MOS and unit effects within the generalizability theory framework.

Prepare draft and final technical report on data analyses. After completing data analysis work as described above, we will prepare a draft report on results of the analyses. The report will be submitted to the COR for review, and we will make the necessary revisions in a final technical report on these analyses.

Subtask 7: Revise Instruments

We will review and as needed, revise both the behavioral rating scales and the objective indices. Specifically, based on performance rating data from the FY83/84 cohort study (first tour), and the data from field testings of the various instruments, revisions to both procedures for collecting objective indices and ratings of soldier effectiveness will be made. Having followed an iterative process throughout the four field tests, the revisions to be made at this time are anticipated to be minor refinements rather than major revisions since by this point the instruments will have already undergone a number of revisions.

To accomplish the objective of this subtask, it is the intention of our staff to coordinate, through the COR, for the review of these instruments by high level Army officials, and then to incorporate changes, as required. Simply stated, this subtask can be viewed as one in which the final "polishing" of instruments will be accomplished prior to administration during the FY86/87 first-tour and FY83/84 second-tour data collection.

Subtask 8: FY83/84 Cohort Second-Tour and FY86/87 Cohort First-Tour Data Collection and Analyses

This subtask consists of three activities: (a) preparation of a draft and final plan for the FY83/84 cohort second-tour and FY86/87 first-tour data collection; (b) administration of the rating and administrative measures to the two samples of cohort members; and (c) analysis of data from these data collections.

Draft and final plan for data collection. The first step in this activity will be the preparation of a Troop Support Request for (a) a sample of the FY86/87 cohort; and (b) those personnel who were measured in the FY83/84 cohort study (Subtask 6) and have remained in the service for a second tour. Of course, attrition will reduce considerably the original sample by the time this second-tour measurement is accomplished. We estimate that only about 10 percent of the original sample will be available for second-tour data collection at the sites visited.

Additionally, a draft plan for data collection from the FY83/84 and FY86/87 cohort will be prepared and submitted to the COR for review. As in Subtask 6, the plan will discuss rating scale administration procedures and collection of objective administrative index data. After review by the COR we will revise the draft plan and submit a final plan for data collection.

Administration of the rating and administrative measures. In this activity, FY83/84 (second-tour) and FY86/87 (first-tour) data will be collected concurrently by the cohort data collection team, i.e., Task 5 researchers, during months 69-72 of the effort. The data collection will proceed in the same manner as has previously been described in Subtask 6. Data on the same objective measures used for the FY83/84 cohort first-tour study, will again be collected for this second-tour cohort, as well as for the FY86/87 first-tour cohort. Additionally, the behavior-based rating scales developed in Subtask 4 to tap second-tour soldier effectiveness will be administered to superiors and peers of the second-tour sample members,

who will complete self-ratings. Finally, supervisory, peer, and self ratings will be collected on the FY86/87 cohort in the same fashion as they were collected on the FY83/84 first-tour cohort (see Subtask 6 for a description of data collection procedures).

<u>Data analyses.</u> The main purpose for gathering these data is of course to evaluate the validity of current and newly developed pre-induction predictors and the in-service predictors developed in Subtask 4. Thus, we will submit the data collected here to Task 1 staff and to the LRDB for the validity analyses. However, as in Subtask 6, the Task 4 researchers will perform analyses to determine the quality of the data collected. As before, these analyses will include but not necessarily be limited to: (a) examining the distributions of ratings and administrative index scores, (b) evaluating the interrater agreement of ratings, (c) examining the dimensionality of the ratings and of the administrative data, and (d) evaluating the convergent and discriminant validity of the ratings and the ratings and administrative indices together.

In addition, as in Subtask 6, we will once again have the opportunity to assess the accuracy of some of the ratings and administrative index scores by correlating Task 4 criterion scores with selected job performance scores from the Task 5 research. However, limitations on interpreting these correlations will be the same here as were discussed previously in relation to first-tour criteria. Likewise, evaluations of MOS and unit effects may proceed within a generalizability theory framework.

Subtask 9: Obtain Scaled Utilities

Much of the work of Project A focuses on the development of new and improved measures of performance. These measures are intended to serve as criteria for use in evaluating alternative selection and classification decisions. The goal of this subtask is to provide the basis for translating the measures of different aspects of a soldier's performance into a single, best indicator of the soldier's net worth to the Army, relative to other soldiers performing at different levels, perhaps even at different tasks.

The primary need for such a measure of each soldier's relative "utility" vs cost to the Army is that a single criterion is needed in implementing CPAS. The decision to accept an applicant into the Army must be based on an estimate of the applicant's potential worth to the Army vs estimated cost in comparison to the potential net worth of other applicants. The decision to classify an applicant into one MOS rather than some other must similarly be based on the applicant's net worth to the Army if assigned to each of the MOS in comparison to other applicants who might be assigned to the various MOS.

In order to compare an applicant's relative worth to the Army in alternative MOS, it is clear that our estimate of worth to the Army must be on a single common scale across the different MOS.

It is further desirable that we be able to express the common utility scale in a "dollar" metric. If each soldier's predicted worth can be expressed in monetary units, it will greatly facilitate cost-benefit analyses of alternative testing and classification procedures. The cost of additional

testing time or of the operation of CPAS and the personnel replacement cost of training, support and maintenance can be traded off against the increases in the dollar worth of the selected applicants.

Finally, in developing an overall scale of utility, it will also be important to state the context in which utility is to be assessed. The importance to the Army of an increment in some type of performance may be quite different during wartime than it is during peacetime, or at different levels of mobilization. We will examine this possibility by collecting utility data under more than one specified context and comparing the results. Where differences are found, results based on the alternative scalings must be passed on to Project B so that CPAS can be adjusted to optimize performance under the different contexts.

While the needs for a common scale of utility are clear, the potential difficulties in the development of such a scale are equally clear. We are attempting to create a common scale of utility across a very wide range of circumstances based on newly developed performance measures. Most examples of successful utility scaling that approach this order of magnitude have been based heavily on relatively "hard" economic data rather than with the more approximate types of performance indicators that can be achieved in this project. If we are even approximately successful in achieving a common dollar-valued utility metric, we will have made a significant advance in the application of utility scaling techniques.

The various problems in the development of the needed utility metric are described below. There are alternative models or procedures that suggest alternative means of handling the potential problems in each case.

Wherever possible, we will seek to try out as many alternatives as are feasible in our testing so that we may obtain some indication of the level of congruence of the results of these alternative approaches. Where there is a high level of agreement between alternative approaches, it may be taken as an indicator of the validity of the approaches considered. Where such congruence is lacking, we must proceed to investigate potential sources of this noncongruence before deciding which approach(es) to use.

Development of performance construct measures. The assessment of each soldiers's worth to the Army must necessarily be based on the performance measures we have identified or developed in Tasks 4 and 5. The nature of these measures varies from indirect indicators of constructs such as "obeys orders" to directly observed measures of the performance of relevant tasks. These measures represent only a sample of the possible measures, e.g., only a sample of the tasks performed within an MOS. However, it is important that this sample adequately represent the entire domain of performance dimensions since they are the only measures we have from which to estimate each soldier's utility.

A problem in the development of an overall utility scale is that the number of individual measures will be quite large. Some reduction in this number is necessary to make the task of identifying the relative importance of each measure more manageable. By this we do not mean that we will seek to identify some abstract set of orthogonal dimensions through factor analytic techniques. Rather, we will build on the multitrait-multimethod analyses already planned for our assessment of construct validity. We will combine the multiple measures (e.g., supervisor and peer ratings, job knowledge and

hands-on performance measures) of a single performance construct into a single best (most reliable) composite based on the LISREL V models used in construct validity analysis. In this discussion, we will call the resulting composites performance construct measures as distinct from observed measures, although if a construct is assessed by only one observed measure, the two will be identical.

For the purposes of obtaining utility judgments, it will be desirable to divide the performance construct measures into distinct levels. The number of levels may range from two, in the case of GO/NO-GO task measures, to as many as five levels for measures such as five-point rating scales. Where several observed measures have been combined, it will be important to label the levels of the composite in terms of the more objective component measures, e.g., at the "superior" level, 85 percent can perform task X correctly on the first try. Once the individual performance constructs and the distinct levels of performance within each construct have been defined for each MOS, we are ready to proceed with the assignment of utilities.

<u>Development of utilities within each MOS</u>. The exact procedure to be used in obtaining utility estimates for performance levels within each MOS will be developed during the first two years of this effort. Because of the unprecedented scope of the present effort, it will be important to conduct tryout studies to test the feasibility and the validity of alternative approaches in the present context. Since the actual data collection for utility scaling does not take place until the latter half of 1985, there will be sufficient time for such tryouts.

In general there are two approaches to the development of utility scales: (a) to use naturally occurring indicators of the relative importance of different levels of performance on the different performance constructs, and (b) to gather judgmental data from qualified respondents. While the former approach may provide some insights, we feel that it is essential to gather judgmental preference data from high ranking military officers. Only in this way can ARI be assured that the selection and classification system will result in personnel assignments that are in accordance with the performance expectations of senior military leaders. We will look for indirect indicators such as differential rates of advancement for soldiers performing at different levels on each construct, but the whole motivation for the effort to develop performance measures stems from the absence of adequate operational indicators of utility.

There are a number of alternative approaches for eliciting utility scaling data from expert judges (see Luce & Suppes, 1965; Raiffa, 1970). These range from approaches identified with Multi-Attribute Utility Theory (see Edwards, 1977; Kenney & Raiffa, 1976) which obtain judgments of the relative importance of different attributes along which the stimuli are rated, to approaches based on a conjoint measurement perspective (Luce & Tukey, 1964) which assess the importance of different attributes through comparisons of individuals with different values on the different attribute (performance construct) dimensions. In the former case, concern is with whether judges are able to give valid ratings of the attribute dimensions directly. In addition, it may be necessary to perform a separate scaling of the utility of each level of each performance construct first before a linear composite function is appropriate. With such independent scalings

of the attribute dimensions, it would not be possible to take interactions among the attributes into account. (The difference between good and mediocre performance on one dimension may depend on the level of some other attribute.) In the case of other approaches, there is still a concern with the validity of the individual judgments, but the primary concern is generally with the number of individual judgments that may be required to yield stable parameter estimates.

Regardless of the exact procedure that is found to be most effective, a vital component in the development of utility scales within each MOS will be the description of the stimuli to be ranked or rated. It is critical that, whether individuals or constructs are being described, the descriptions be behaviorally anchored to the maximum extent possible. Rather than saying that a soldier to be rated had "moderate" discipline problems, for example, we would say that this soldier had two disciplinary reports in the past year. Rather than saying "performs task X well", we prefer statements such as "85 percent of the soldiers at this level can perform task X correctly on the first try" (or "within 15 minutes").

In collecting within-MOS utility scaling data, we will first focus on the nine MOS for which hands-on performance data will be available. For each of these nine MOS, we will develop procedures and stimuli and try them out, first with project and ARI staff, and then with three SME. The procedures will be tried twice in counterbalanced order: (a) with performance constructs defined in terms of all the measures available for that MOS, and (b) with the constructs defined in terms of only the Army-wide measures

available for the MOS. Insofar as possible, microcomputers will be used to control the presentation of stimuli and the collection of responses.

Following the procedures outlined in the next section, we will then rescale the alternate sets of composite utilities for the nine MOS onto a common scale, again using small groups of raters performing the task in counter balanced order. Based upon the scale values obtained, we will compute two sets of composite utility scores for the soldiers tested[1] in each MOS. If (as we suspect) these utilities correlate quite highly and have equivalent means and variances within MOS and produce the same order of statistical differences across MOS, it will then be reasonable to use the performance constructs defined in terms of only the Army-wide measures when we rescale the utility scales for all sampled MOS onto a common scale. If the two sets of composite utility scores for the tested soldiers are quite different, it will mean that the inclusion of the hands-on performance levels in the definitions of the constructs substantially influenced the raters. Utilities derived exclusively from Army-wide measures for the nine MOS for which we do not have hands-on measures will then have to be interpreted cautiously if some way to adjust the utilities through statistical or judgmental means is not found.

When the exploratory research with the first nine MOS is completed, we will turn to the ten MOS for which only Army-wide measures are available. The

---

[1]Initially, data from the field trials of the performance measures, or even dummy data, will be used. When the data from the FY83/84 cohort become available, these analyses will be repeated.

process yielding the most equivalence between the two sets of utilities in the first nine MOS will be repeated using the available measures. When we are satisfied with our approach we will obtain within MOS utility scales for all 19 MOS using larger groups of SME.

Rescaling the utility of each MOS onto a common scale. While it may be possible to develop a scale that is constant across different MOS at the same time that the within scales are developed, we do not expect to do so. The division of the problem into separate steps for within and across MOS utilities allows us to examine the results and assumptions of the more detailed within MOS scaling before proceeding to the next stage. It is almost surely the case that the number of judgments required to perform the entire scaling task in one step would be prohibitive. In addition, we think it is likely that the within-MOS utility scaling would benefit from the advice of raters with detailed knowledge of the MOS being rated, while the across-MOS utility scaling requires judgments from officers at a relatively higher level of command.

Because of the complexity of the evaluation, it would be unwise to expect raters to be able to holistically order the importance of performance increments across the different MOS in the absence of more specific criteria for evaluation. Our approach, therefore, will be to use an iterative procedure, involving consultation with relevant literature and expert opinion, to distill a good set of general evaluative criteria (Army goals) that are broad in scope yet meaningful, practical, and internally consistent. Some of the factors and issues that will be considered in formulating an attribute set for evaluating MOS utility include: (a) impact on force readiness,

on military survival; (b) centrality of MOS activity; (c) contribution to psychological and physical well-being of forces, and to civilian relationships and support; (d) effects on enemy performance; etc. These types of factors and numerous others will be itemized, and then configured into a coherent set of evaluative criteria.

The criteria that will be used to assist raters in judging the relative importance of performance increments in different MOS will be determined by conducting half-day workshops with two groups of 10-15 more senior military officers. The participants in each workshop will carry out two tasks. First, they will use critical incident methodology to generate specific examples of when performance in a particular MOS was judged to be particularly effective or ineffective. They will also be asked to list the reasons why the particular episode was judged to be effective or ineffective. Second, the participants will be presented with a series of choices to be made under various contexts such as the following:

> Suppose during wartime you have ten Wheel Vehicle Repairers (63W) and ten Track Vehicle Repairers (63Y). Five of the 63W and five of the 63Y are highly competent and the others make frequent mistakes. If you could replace two of the less able Repairers for the more able ones, would you choose to do this for the 63W or the 63Y?
>
> (If the judge favors the 63W:) How many 63Y replacements would be of equal value to the replacement of two 63W?

After his or her choices are made, the participant will be asked to verbalize, or write down, the reasons for the choice.

On the basis of the content of the reasons given for the choice of critical incidents and the reasons for the hypothetical choices, the research staff will derive the most parsimonious and most frequently cited components of importance for each context separately and combined. The content analysis will be done separately for the two workshops, and only components, i.e., reasons, upon which the two groups agree will be used.

At the level of across MOS comparisons, it will be desirable to identify four or five levels of performance on the within-MOS utility scales and to again anchor these levels in descriptions of the constructs (maybe half a dozen or so) most relevant to the utility scale for that MOS. Thus a "superior infantryman" might be defined in terms such as:

(1) 95 percent of the infantrymen at this level can fire a rifle within some specified level of accuracy,

(2) 90 percent of the infantrymen at this level can clean and reload rifle correctly within x minutes, and

(3) fewer than 5 percent of the infantrymen at this level have had any discipline problems at all during the past year.

We expect something on the order of five to ten statements to "define" each performance level for the MOS. We will investigate the feasibility and validity of alternative procedures for eliciting the required ratings from expert judges. One of the methods that will be explored is to ask first for ratings of the relative importance of the different evaluation criteria to the overall mission of the Army. The judges would then be presented with descriptions of different performance levels in different MOS and asked to rate the value of various performance increments according to each

criterion. Then overall evaluations of performance increments in the different MOS can be computed and displayed to the judges. The judges would be allowed to modify their ratings on the basis of this feedback.

For both the within-MOS and the across-MOS utility scaling, it will be important to elicit a sufficient number of judgments to significantly overdetermine the appropriate scaling. In this way it will be possible to create internal evidence of the consistency with which judgments have been made, both within and across raters. As in the case of the within MOS scaling, we will proceed from tryouts with small groups of raters to larger groups in order to obtain more stable and valid values using the procedures found to be most appropriate.

<u>Assigning dollar values to performance utility levels</u>. As stated above, it is highly desirable that a translation of the utility scale into dollars be derived. This step is essential in comparing the benefits (in the utility of increased performances) of CPAS to the related costs. One such comparison of particular interest is between the costs and benefits of various increases in testing time. Fortunately, it will not be necessary to obtain direct dollar values for all MOS/performance level combinations if the performance utilities have been reliably and validly scaled. A sample of MOS/performance levels can be monetized and the dollar values of the utilities of the remaining combinations obtained through derived utility/dollar translation curves.

The problems of achieving the desired translation, however, are many and large. Short of detailed simulations whose cost would be prohibitive, we

are not optimistic about achieving more than a very approximate translation. We will, however, allocate time for further reviews of relevant literature and for discussions with others who are trying to derive similar or related estimates. At present, we see three general approaches to this issue. We will explore the feasibility of obtaining a translation through each approach so that we can use the level of agreement as an indicator of the accuracy of the final translation. The three approaches are:

(1) assessment of the value of performance through surveys of decision-makers,

(2) inference of the value from cost data and some assumptions about the relation between cost and value, and

(3) comparisons with the civilian sector.

Sinden and Worrell (1979) describe several methods of eliciting valuation judgments. "Direct questioning" approaches could be used to ask military officials how much more they would be willing to pay for specified increments of performance. Other approaches involve "budget allocation" or different "trade-off games." These latter approaches, which are favored by the authors, could involve comparisons of performance increments with hardware or other items with known dollar costs. We will develop and try out alternative variants of these approaches and administer the best of them at the same time that the across MOS judgments are collected.

In exploring the second approach, the proposed methodology is based on the economic concept of the duality of cost and output functions. (A profit maximization problem subject to a cost constraint has a "dual" which is a cost minimization problem subject to a production constraint.) Relative

values of inputs may be inferred either from production data or from cost data. In the present case, the valuation of performance might be inferred from estimates of the cost of producing a given level of performance.

If this project were being conducted as a profit maximizing business firm, an obvious candidate for use in valuing performance in various occupations would be wages (adjusted to include fringe benefits and associated costs of employment borne by the employer). Such costs would represent the cost to the employer of maintaining a worker in a given job. Most economists would deem this approach superior to surveying managers and asking for relative valuation of output from employees at differing performance levels, even though there would be some variation in performance within a wage class.

In the present case, however, the Army does not pay in proportion to productivity or opportunity cost of labor except in relatively indirect ways such as reenlistment bonuses. Instead, wages are, in effect "administered prices" and, as such, are inappropriate for determining the value of performance differences directly. An alternative concept that takes account of all costs associated with Army personnel at given performance levels is <u>replacement cost</u>. This cost includes the expected recruitment, testing, processing, training, and compensation costs necessary to replace a soldier functioning at a given level of performance in a given MOS.

It is important to note that, as with other settings in which administered prices exist, shortages and excess supplies may result. (For a recent review of some of the effects of administered prices, see Jacob Mincer, 1982). These results impose other costs on the Army which must be considered as part of the total cost of maintaining a soldier in a particular MOS

at a particular performance level. For example, in occupations for which Army pay is less than alternatives in the civilian sector, recruiting cost to fill such jobs may be high and loss rates due to failures to reenlist may also be high. On the other hand, if Army pay is better than pay in civilian jobs, there may be excess supply to the Army resulting in low recruiting costs and high reenlistment rates. Because replacement cost includes attrition and non-reenlistment losses, it is comprehensive enough to take account of loss rates attributable to non-comparability of civilian and military pay scales. (Because training is costed explicitly there is not a problem that might exist with civilian wages.)

For any MOS/job performance category, the cost of replacing an individual soldier may be computed. Cost data on training, testing, and processing should be directly available from the Army MIS, as are separation rates; however, recruiting costs by recruit "quality" levels will have to be inferred. Some econometric studies of the determinants of recruit supply can be used to infer the marginal cost of recruiting by education and test score category (see Dale & Gilroy, 1983 and Huck & Midlam, 1977, for examples).

The third approach will be to examine data on workers employed in the civilian sector. Wage differentials could be related to differences in either ratings of performance levels or performance levels predicted from employee ability and aptitude. If both ability and performance measures correlated poorly with wages for specific occupations, this would suggest that cost utility might not be sensitive to performance level for that

occupation. This approach will require finding comparable civilian occupations for at least some of the MOS studied and using longitudinal data sets that contain both wage and ability estimates for these occupations. One example of such a database is the Project TALENT study (Wise, McLaughlin, & Steel, 1978) which contains information on earnings at age 29 of over one hundred thousand individuals in civilian occupations and also scores from the prior administration of a two-day test and questionnaire battery.

In addition, research which relates cognitive and non-cognitive abilities to wages, supervisor ratings and other evaluative measures of performance in the civilian sector will be reviewed for insight into relationships betweenperformance and wages. (Examples include Gintis, 1971; Grilich & Mason, 1972; and Wise, 1975.)

Schedule and troop support requirements. This subtask will be accomplished jointly by Task 4 and Task 1 personnel, with Task 4 personnel having primary responsibility for the collection of the data and Task 1 personnel primary responsibility for the data analysis. As shown in the Integrated Master Plan, the subtask has been divided into three periods of activity. During the development stage (December, 1984 through July, 1985), we will first require the assistance of three subject matter experts in each of the nine MOS for which MOS-specific measures are developed. Three interactions of roughly one-half day each are anticipated. In addition, we will need to select and begin to work with a sample of 30 military officials who will supply the initial ratings of utility estimation. Each will participate in one of the half-day workshops to develop the criteria to aid the across MOS judgments, and on a selective basis, on the tryouts of the judgmental

methods. When the preliminary research on the nine MOS has been accomplished, additional SME will be used to scale the performance constructs within the MOS for which only Army-wide measures are available. The military officials will then scale the MOS/performance level combinations for all 19 MOS.

During the second stage (August, 1985 through February, 1986), we will examine the score distributions, reliability, and intercorrelation of the performance measures obtained for each MOS and determine whether the within-MOS performance construct scaling should be redone (as may be the case if we decide to drop some measures or combine others in different ways based on the empirical results). If so, an additional one-half day will be required from three SME in each MOS to provide data for within MOS scalings. In either event, one additional day will be required from each of 30 senior experienced officers (a new sample) to provide the cross-MOS ratings at each level (including cost/value estimation). These sessions can be scheduled independently at the convenience of the officers involved. The utility values obtained will be used to help evaluate the cost effectiveness of various measurement alternatives in Task 2 decisions concerning the composition of the Experimental Battery.

During the final stage (August, 1988 through March 1989), the data collection and analysis process will be repeated to allow for the incorporation of second-tour measures (administered to the FY83/84 cohort) and to accommodate any changes in the performance battery used in the longitudinal validation (FY86/87 cohort). In addition, the analysis will examine the

consistency between the utilities obtained earlier (in 1986) and the more recent set in an attempt to identify the impact on the utilities of such factors as inflation, changes in the civilian labor market and the U.S. general military stance, and innovations in military doctrine, equipment, and manning policies. The troop support requirements will be the same as for the second stage.

Subtask 10: Prepare Final Reports

We will first prepare a draft final report describing all Task 4 research to this point. The report will be submitted to the COR and revisions made based on his feedback. We will then submit the final report incorporating all comments and suggestions from the COR.

SUMMARY OF EXPECTED OUTCOMES FROM TASK 4

Operational Outcomes

The following outcomes of the research should be useful to the Army from an operational standpoint:

1. The model of soldier effectiveness should provide a concrete, behavioral definition of what is expected of a first-tour soldier in the Army. This definition can in turn be used to explain to a first-tour enlisted person what is expected of him/her. The model could be made part of indoctrination courses, for example, to demonstrate to incoming troops what they should be striving for as soldiers in the U.S. Army. We see great advantages to using the model to help indoctrinate soldiers to the Army--for one thing, the behavioral nature of this model will make unambiguous the communicated expectations for effective soldier performance. A second point is that presenting a single model to all incoming enlisted personnel provides a common set of expectations so that everyone gets the same information about these expectations, and misleading guidance on soldier performance requirements is avoided.

2. Another possible application of the soldier effectiveness model is to provide standardized guidance to recruiters about what is expected of first-tour soldiers so that they can pass this information on to prospects and recruits. The model's dimensions and behavioral performance requirements can be packaged to present a realistic but also highly motivating depiction of what it takes to be a successful

soldier, and recruiters can be instructed on how to present these materials to prospects/recruits. The package on the model might be used then to help sell prospects on an Army enlistment by essentially providing an in-depth definition of what is meant by "soldiering" in the Army, and also as a kind of realistic job preview (Wanous, 1973) to show recruits what performance requirements to expect during their first term.

3. A third use for the model of soldier effectiveness is to provide new dimensions for the EER. The model's dimensions should be ideal as EER dimensions because they will be based on an across-MOS, Army-wide analysis of enlisted jobs, and the performance requirements emerging in the model will reflect concrete, observable, dimensions of soldier effectiveness.

4. The rating scale administration package and procedures can be used in future personnel research in the Army. A major effort in the present research will be to develop an effective but very efficient set of procedures for administering performance rating scales to large numbers of persons. These procedures and the package of materials found most effective and efficient can certainly be adapted for use in other Army personnel research where ratings of many persons are required.

5. Likewise, we will develop in the research a system and procedure for scoring individual enlisted personnel on administrative index composites and on attrition/reenlistment categories. Future personnel research requiring soldier effectiveness scores can certainly use this system/procedure.

6. Also likely to arise from Task 4 research are guidelines to record and collect administrative data more consistently across units. These guidelines should increase considerably the quality and usefulness of the administrative data for indexing aspects of soldier effectiveness. Thus, in future personnel research efforts requiring such effectiveness scores on enlisted personnel, correspondence of these scores across units and the accuracy of the data should be enhanced substantially over the scores on administrative indices presently available.

Scientific Outcomes:

The following scientific outcomes are anticipated from the Task 4 research:

1. A major theme of the research in Task 4 is the evolving of principles and conclusions regarding effective and efficient methods of gathering rating data from large numbers of persons. What we learn in this effort, e.g., the kinds of instructions to raters, rules for selecting raters, training and orientation for raters, etc., that lead to relatively high quality ratings, will be important knowledge that can be applied in all situations where large-scale rating data collection is required. Considerable research has been done on the effects of different rating formats on quality of ratings (e.g., Dunnette & Borman, 1979; Landy & Farr, 1980), but work is needed to specify aspects of the context in which raters are placed that likewise influence quality of ratings. The series of field tests planned for Task 4 research should yield considerable knowledge in this area.

2. A trend in the recent literature has been to examine accuracy of ratings (where possible), rather than or in addition to assessing psychometric characteristics of ratings such as halo, leniency, and restriction of range (Bernardin & Pence, 1980; Borman, 1979). The argument is that these two sets of rating quality criteria often do not correspond very closely, and on a conceptual basis accuracy is definitely the criterion of most importance. As an example, we have found that a certain rater training program reduced halo but left accuracy of the ratings unaffected (Borman, 1979).

Although we look favorably on the trend toward considering accuracy (e.g., Dunnette & Borman, 1979), a criticism of research done on accuracy of ratings is that it has been performed in the laboratory using "paper-people" ratees (stories about how someone performs on a job) or videotaped performers to be rated (Campbell, 1978). These settings for accuracy research may not be the most realistic, and research findings in the laboratory might not generalize well to organizational settings.

A problem here is that typically, organizational settings provide no opportunity to obtain actual or "true" performance scores on individuals against which to compare ratings of their performance. For the vast majority of jobs no absolute standards exist to enable assignment of true performance scores to employees, and thus, accuracy of ratings cannot be evaluated.

All of this leads to the observation that in this research we will have available at least an approximation of these true scores. Task 5 researchers will develop performance tests presumably yielding high fidelity performance scores, providing an accurate picture of each soldier's actual performance level (at least in the technical competence aspects of the job). We anticipate using these performance scores as standards against which to compute the accuracy of Task 4 ratings made in the research.[2]

We .ery enthusiastic about this opportunity to bring research on performance rating accuracy out into an actual organizational setting. Examining the effects of different rating formats, various administrative sets in which raters are placed, different rater training and orientation procedures, etc. on rating accuracy in a "real" organizational context should yield very important results and conclusions bearing on how to generate more accurate ratings.

3. Some research has examined supervisory, peer, and self-rating sources comparing the relationships between ratings sources (e.g., Heneman, 1974; Klimoski & London, 1974) and the contributions to validity of

2Performance measurement in Task 5 relates to the "can-do" part of the criterion space, that is, the skill or ability-related proficiency aspects of performance. The "will-do" criterion space, that is, the soldier's performance over time on the job, his/her continuing motivation to succeed, etc., cannot be well tapped by this kind of performance measurement. Thus, we must be careful correlating Task 4 ratings with Task 5 performance scores and how we interpret these relationships. However, Task 4 ratings of a soldier's technical competence, for example, can be justifiably compared to Task 5 performance scores to evaluate the accuracy of Task 4 ratings.

ratings from these sources (Borman, 1974; Buckner, 1959; Campbell, Dunnette, Lawler, & Weick, 1971). Typically, however, studies take place in a single organizational setting, with relationships between raters, i.e., supervisors and peers, and ratees fixed for the most part. In other words, held constant is the familiarity of the ratees' work on the part of each organizational level's raters. For example, peers may be very familiar, and supervisors not at all familiar with a ratee's work in an organization studied. In addition, raters who come from different organizational levels are likely to view a ratee's performance within a narrowly defined set. For example, supervisors may be able to view only a ratee's interpersonal skills, while peers may view a ratee's behavior in all aspects of the job.

The point is that the Army offers great variety in this regard. Some units, e.g., infantry, have both supervisors and peers who will be very knowledgeable about ratee performance in all aspects of the job. Other units, e.g., maintenance, may have raters from different organizational levels viewing very different aspects of a ratee's job performance. Therefore, research can be conducted to evaluate how different rater-ratee work relationships and different opportunities on the part of the raters to view ratee performance influences their performance/effectiveness evaluations of ratees. An obvious question here is how these relationships and opportunities relate to interrater agreement in ratings and to accuracy of the ratings.

4. Research on composites of low base rate objective measures may lead to some general guidelines for how base rate problems can be dealt with in

personnel research. Low base rates in psychological research have been an acknowledged difficulty for many years (Meehl & Rosen, 1955), and forming composites of them is one possible approach to alleviating these problems. Task 4 research can assess the usefulness of this approach.

5. James (1973), Smith (1976), and others have written about construct validity principles applied to criterion development, but little has been done to follow up with such applications. Likewise, in our proposal we discuss working on Army-wide criterion development steps within a construct validation framework. However, in the Task 4 research we plan to <u>put into practice</u> these construct validity principles.

First, the model of soldier effectiveness is meant to be an inductively derived behavioral definition of the dimensions of Army-wide performance and effectiveness. This behavioral definition should exhaust the domain of important performance requirements and effectiveness dimensions that pertain to all MOS, and will drive development of measures to tap each element of the model. In the measure development work, careful attention will be directed to selecting the most appropriate method(s) to index performance/effectiveness for each of the model's dimensions. Supervisor and peer ratings, along with objective indices of effective and ineffective soldier behavior, will be targeted toward the appropriate model dimensions.

Regarding analysis of criterion data, two conceptions will guide our efforts. First, we will use ideas of multitrait-multimethod analysis and convergent and discriminant validity of measures (Campbell & Fiske, 1959; Kavanagh, MacKinney, & Wolins, 1971) to evaluate the quality of objective indices and ratings from different sources. Wherever possible we will strive for convergence across methods in measuring a performance/effectiveness construct and also for <u>differentiation</u> in measurement of very different kinds of constructs.

The reason differentiation is important is that prediction-criterion links tend to make better conceptual sense when specific rather than global criteria are available. This is the same reasoning used in developing categories of attrition. Very different predictors are likely to be appropriate in predicting medical attrition and attrition for disciplinary reasons.

To aid in differentiating between constructs in measuring criterion performance and effectiveness, we will apply a strategy that combines factor analysis of criterion data with hypotheses about underlying performance/effectiveness constructs. These hypothesized constructs should reflect important, homogeneous variables within constructs, and yet, they should show conceptually different content across the constructs.

As mentioned previously, we have some hope that such an approach will be fruitful. Borman, Rosse, and Abrahams (1980) discovered that a conceptually reasonable 3-factor structure described Navy recruiter

performance, and this 3-dimension solution was replicated in two other samples. Also, in ratings of transmission and distribution, worker performance, a 2-factor solution, made excellent conceptual sense (technical competence and interpersonal adjustment to job demands; Borman, Mendel, Lammlein, & Rosse, 1981). Thus, in Task 4 research, we will be alert to possible underlying constructs that might map the performance/effectiveness domain in a conceptually appropriate manner, and which can be differentially measured with our criterion instruments.

In sum, we believe that reliance on these and other applications of construct validation thinking will lead to: (a) a better conceptual picture of the Army-wide criterion domains; (b) more accurate measures of the relevant performance/effectiveness constructs; and (c) more meaningful and valid predictor-criterion relationships.

6. An intention with the model of soldier effectiveness is to define a broad set of domains relevant to a soldier's worth or value to his/her unit and the Army. This is a broader view of performance/effectiveness than is typically focused upon, and we may learn something new from taking this approach. For example, the expanded conception of soldier effectiveness may better address some of the intersections between individual and organizational effectiveness. Consider the organizational commitment domain introduced in the preliminary model of soldier effectiveness. This has nothing to do with individual performance as we usually think about it, but commitment on the part of some critical percentage of a unit's members might have considerable impact on the unit's overall effectiveness.

We do not mean to say that this contract will solve many problems in the area of organizational effectiveness. However, the relatively broad view of soldier effectiveness may help to shed some light on how individual performance/effectiveness relates to organizational effectiveness. Results of this view will also be interesting in their own right. We have indicated some possible domains that might emerge from model development work, but others may be identified, as well, in this work. Thus, the content itself of the total set of dimensions developed in the model will be of interest.

7. Although the work is admittedly exploratory, we should learn something about performance ratings from development and testing of the Combat Performance Prediction Scales. What is planned here is very unusual in relation to performance rating formats and, especially, to what is required of raters. Format development steps will lead to dimensions of performance reflecting a completely different context, i.e., combat, from the one currently being experienced, i.e., garrison setting. Thus, it will be of interest to see how scale development works within these contraints.

More scientifically compelling, however, is learning about the effect on ratings of the unusual task asked of raters using the scales, i.e., evaluate how <u>you believe</u> each soldier would perform in a very different setting. The rater must observe and recall the soldier's behavior in the garrison/field setting and make inferences about how he would perform in combat. This is somewhat akin to making "ratings of potential" for higher level jobs, as is done in many organizations. However, it

is likely that the jobs for which predictions are being made are typically more similar in context to the present job than is the case when combat performance is being predicted based on garrison performance.

It may be that raters will have considerable difficulty making these inferences. The research planned will explore how these inferences are attempted, whether or not raters can differentiate between present and predicted performance, and what levels of interrater agreement emerge in such a rating task. These and possibly other analyses should tell us something about the evaluative judgment process under conditions that require considerable inference.

8. Certain results from Task 4 research should bear on the "trait-situation controversy" in personality psychology. Briefly, the trait side argues that relatively stable personal characteristics in individuals for the most part determine behavior in a variety of different situations (e.g., Block, 1971). "Situationists" argue that characteristics of the context or the situation in which people find themselves largely dictate behavior, no matter what the personal characteristics are of the individuals involved (e.g., Shweder, 1975). There is also an interactionist position that considers behavior to be a function of an interaction between the person and situations (e.g., Bowers, 1973). The issues in this controversy are complex and technical, but this is sufficient for our purposes.

Within the present research, consider second-tour performance. It would seem to be a function of: (a) personal characteristics

- those that persons bring with them to the service; and (b) first-tour experiences - characteristics of the unit, quality of training and leadership experienced, etc. These can be viewed as trait versus situation factors and we will be able to get a general idea of the contribution of each to second-tour performance. This is because test and inventory scores will be available for many soldiers (an index of traits). Thus, by comparing correlations between inventory scores and second-tour performance to correlations between first-tour experience responses and that performance, we can gain some idea of the relative contributions of traits versus the situation to second-tour soldier effectiveness.

9. Finally, the Task 4 research program will produce recommendations on practical procedures for determining performance utilities in complex employment situations involving a number of different jobs and settings. We will develop a computer-administered standard tool that can be used repeatedly to derive, extend, modify and maintain utilities. The availability of a relatively easy-to-follow procedure for measuring utility will promote the use of formal decision rules in selection and classification research and in employment, thus increasing the efficiency of personnel utilization in our society.

In sum, we believe that the Task 4 research program should produce both significant scientific and operational outcomes. That is, the expected outcomes of Task 4 will contribute to our general knowledge and understanding of performance/effectiveness measurement in a large organization and should also contribute to the operational needs of the U.S. Army.

REFERENCES

Bernardin, H.J., & Pence, E.C. Effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 1980, 65, 60-66.

Block, J. Lives through time. Berkeley, CA: Bancroft Books, 1971.

Borman, W.C. The rating of individuals in organizations: An alternate approach. Organizational Behavior and Human Performance, 1974, 12, 105-124.

Borman, W.C. Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 1975, 60, 556-560.

Borman, W.C. Format and training effects on rating accuracy and rater errors. Journal of Applied Psychology, 1979, 64, 410-421.

Borman, W.C., & Dunnette, M.D. Behavior-based versus trait-oriented performance ratings: An empirical study. Journal of Applied Psychology, 1975, 60, 561-565.

Borman, W.C., & Peterson, N.G. Selection and training of personnel. In G. Salvendy (Ed.), Handbook of Industrial Engineering. New York: John Wiley, in press.

Borman, W.C., Hough, L.M., & Dunnette, M.D. Development of behaviorally based rating scales for evaluating the performance of U.S. Navy recruiters. Navy Personnel Research and Development Center Technical Report TR-76-31, 1976.

Borman, W.C., Johnson, P.D., Motowidlo, S.J., & Dunnette, M.D. Measuring motivation, morale and job satisfaction in Army careers. Minneapolis: Personnel Decisions, Inc., 1976.

Borman, W.C., Mendel, R.M., Lammlein, S.E., & Rosse, R.L. Developing and evaluating the validity of a test battery to predict performance in transmission and distribution jobs at Florida Power and Light. Minneapolis: Personnel Decisions Research Institute, 1981.

Bowers, K.S. Situationism in psychology: An analysis and a critique. Psychological Review, 1973, 80, 307-336.

Buckner, D.N. The predictability of ratings as a function of interrater agreement. Journal of Applied Psychology, 1959, 43, 60-64.

Campbell, D.T., & Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Campbell, J.P. What we are about: An inquiry into the self concept of industrial and organizational psychology. Presidential address to Division 14, American Psychological Association, Toronto, August, 1978.

Campbell, J.P., Dunnette, M.D., Arvey, R., & Hellervik, L. The development and evaluation of behaviorally based rating scales. Journal of Applied Psychology, 1973, 57, 15-22.

Campbell, J.P., Dunnette, M.D., Lawler, E.E., III, & Weick, K.E., Jr. Managerial behavior, performance, and effectiveness. New York: McGraw-Hill, 1970.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Dale, C., & Gilroy, C. The effects of the business cycle on the size and composition of the U.S. Army. Atlantic Economic Journal, 1983, 11.

Dunnette, M.D. Personnel selection and placement. Belmont, CA: Wadsworth, 1966.

Dunnette, M.D., & Borman, W.C. Personnel selection and classification systems. In L.W. Porter and M.R. Rosenzweig (Eds.), Annual Review of Psychology, 1979, 30, 477-525.

Edwards, W. How to use multiattribute utility measurement for social decision making. IEEE Transactions on Systems, Man, and Cybernetics, 1977, SMC 7, 326-340.

Flanagan, J.C. The critical incident technique. Psychological Bulletin, 1954, 51, 325-58.

Gintis, H. Educational technology and the characteristics of worker productivity. American Economic Review, 1971, 61, 266-279.

Griliches, Z., & Mason, W. Education, income, and ability. Journal of Political Economy, 1972, 80, 74-103.

Guion, R.M. Personnel testing. New York: McGraw-Hill, 1965.

Hammer, T.H., & Landau, J. Methodological issues in the use of absence data. Journal of Applied Psychology, 1981, 66, 574-581.

Hamner, W.C., Kim, J.S., Baird, L., & Bigoness, W.J. Race and sex as determinants of ratings by potential employers in a simulated work sampling task. Journal of Applied Psychology, 1974, 59, 705-711.

Hebein, J., Kaplan, A., Olmstead, J.A., & Sharon, B. NCO Leadership: Task, Skills, and Functions. Alexandria, VA: Human Resources Research Organization, Final Report, February, 1983.

Heneman, H.G., III. Comparisons of self- and superior ratings of managerial performance. Journal of Applied Psychology, 1974, 59, 638-642.

Huck, D., & Midlam, K. Development of methods for analysis of the cost of enlisted attrition. General Research Corporation Technical Report, 1977.

James, L.R. Criterion models and construct validity for criteria. Psychological Bulletin, 1973, 80, 75-83.

Kavanaugh, M.J., MacKinney, A.C., & Wolins, L. Issues of managerial performance: Multitrait-multimethod analyses of ratings. Psychological Bulletin, 1971, 75, 34-49.

Keeny, R., & Raiffa, H. Decisions with multiple objectives. New York, NY: Wiley-Interscience, 1976.

Klimoski, R.J., & London, M. Role of the rater in performance appraisal. Journal of Applied Psychology, 1974, 59, 445-451.

Landy, F.J., & Farr, J. Performance rating. Psychological Bulletin, 1980, 87, 72-107.

Landy, F.J., & Trumbo, D.A. Psychology of work behavior. Homewood, IL: Dorsey, 1980.

Lawler, E.E., III. The multi-trait-multi-rater approach to content validity. Journal of Applied Psychology, 1967, 51, 369-381.

Luce, R.D., & Suppes, P. Preference, utility, and subjective probability. In Luce, et.al. (Eds.), Handbook of Mathematical Psychology, III. New York, NY: Wiley-Interscience, 1965.

Luce, R.D., & Tukey, J.W. Simultaneous conjoint measurement: A new type of fundamental measurement. Journal of Mathematical Psychology, 1964, 1, 1-27.

Meehl, P.E., & Rosen, A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. Psychological Bulletin, 1955, 52, 194-216.

Mincer, J. The Economics of Wage Floors. New York, NY: Columbia University, 1982.

Motowidlo, S.J., & Borman, W.C. Behaviorally anchored scales for measuring morale in military units. Journal of Applied Psychology, 1977, 62, 177-183.

Peterson, N.G., Houston, J.S., & Rosse, R.L. The LOMA job-effectiveness prediction system: Job performance criteria. Submitted to Life Office Management Association, July, 1981.

Raiffa, H. Decision analysis: Introductory lectures on choices under uncertainty. Reading, MA: Addison-Wesley, 1970.

Schmidt, F.L., Hunter, J., & Pearlman, K. Assessing the economic impact of personnel programs on work force productivity. Personnel Psychology, 1982, 35, 333-347.

Schmidt, F.L., Hunter, J.E., & Urry, V.W. Statistical power in criterion-related validity studies. Journal of Applied Psychology, 1976, 61, 473-485.

Shields, J.L., Hanser, L.M., Williams, E.W., & Popelka, B.A. Pilot research for validation of ASVAB and enlistment standards against performance on the job. Paper given at Military Testing Association, October, 1981.

Shweder, R.A. How relevant is an individual difference theory of personality? Journal of Personality, 1975, 43, 455-484.

Sinden, J.A., & Worrell, A.C. Unpriced values: Decisions without market prices. New York, NY: Wiley-Interscience, 1979.

Smith, P.C. Behaviors, results, and organization effectiveness: The problem of criteria. In M.D. Dunnette (Ed.), Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally, 1976.

Smith, P.C., & Kendall, L.M. Retranslation of expectations: An approach to the construction of unambiguous anchors for ratings scales. Journal of Applied Psychology, 1963, 47, 149-155.

Steers, R.M. Antecedents and outcomes of organizational commitment. Administrative Science Quarterly, 1977, 22, 46-56.

Taylor, E.K., & Wherry, R.J. A study of leniency in two rating systems. Personnel Psychology, 1951, 4, 39-47.

Terborg, J.R., & Ilgen, D.R. A theoretical approach to sex-discrimination in traditionally masculine occupations. Organizational Behavior and Human Performance, 1975, 13, 352-376.

VanMaanen, J., & Schein, E.H. Toward a theory of organizational socialization. In B.M. Staw (Ed.), Research in organizational behaviors (Volume 1). Greenwich, CT: JAI Press, 1979.

Wanous, J.P. Effects of a realistic job preview on job acceptance, job attitudes, and job survival. Journal of Applied Psychology, 1973, 58, 327-332.

Wernimont, P.F., & Campbell, J.P. Signs, samples, and criteria. Journal of Applied Psychology, 1968, 52, 372-76.

Wise, D. Academic achievement and job performance. American Economic Review, 1975, 65(3), 350-66.

Wise, L.L., McLaughlin, D.H., & Steel, L. The Project TALENT Databank Handbook. Palo Alto, CA: American Institutes for Research, 1978.

TASK 5 RESEARCH PLAN

MEASUREMENT OF MOS-SPECIFIC PERFORMANCE

GENERAL PURPOSE OF TASK 5

The type of linked personnel decision-making system that will result from this project has long been of interest to the Army. However, the Army currently has neither the system nor the data to make critical personnel decisions throughout a soldier's life cycle based on the soldier's job performance and the needs of the Army. In the system currently in use, the initial selection and classification decisions are predicated on the relationships of entrance tests to performance in the Advanced Individual Training (AIT) environment. They are not tied to performance on the job. In fact, with few exceptions, entrance tests have not been validated using job performance as a criterion.

Before we can evaluate the relationships between scores on predictor instruments and actual job performance we must resolve the criterion problem, the key problem in this project, and the joint objective of Tasks 4 and 5. Task 4 is concerned with the development of valid measures of overall performance as a soldier; i.e., of constructs that apply to all MOS. The purpose of Task 5 is to develop criterion instruments that accurately measure MOS-specific job performance.

BACKGROUND ISSUES AND RATIONALE

Developments in two areas set the stage for work on Task 5: performance evaluation in the Army and job and task analysis.

Performance Evaluation in the Army

The Army's mobilization experience in World Wars I and II led to a classification process that was based on matching an individual's civilian job skills with those of a comparable military job. The process emphasized occupational code equivalencies rather than independent standards of job competency. It was not until 1955 when the Army instituted the Enlisted Evaluation System (EES) that standards were established and proficiency began to be assessed. The EES used a job knowledge test--more commonly referred to as the MOS test or Pro Pay Test--together with a Commander's Evaluation Report (CER) in annually evaluating enlisted personnel in grades E-3 and above. Results of the evaluations were used chiefly for personnel management purposes: to determine eligibility for reenlistment, promotion, proficiency pay, additional schooling and the like.

The MOS test was a norm-referenced achievement test designed to measure, in a broad sense, job knowledge. No claim was made that the MOS test measured job proficiency, although by implication something related to job proficiency was being tapped. Standard four-alternative multiple-choice test questions were drafted by military SME personnel in Item Writing Agencies (service schools). The questions were based on test outlines prepared by the Enlisted Evaluation Center (EEC) from functional MOS descriptions

contained in AR 611-201. Thus the content of a test tended to reflect the MOS-producing training program; its content validity was limited to the degree to which the field requirements of an MOS were reflected in the MOS functional description and training program. Psychometricians, who staffed EEC, edited and revised items submitted by the schools, assembled the 125-item tests, sent them to the field for administration by a Test Control Officer, scored the returned tests and combined each score with the commander's twelve-factor rating (CER) in reporting to each soldier his MOS Evaluation Score.

EEC maintained item banks based on the conventional internal consistency item statistics. With isolated exceptions, the MOS tests did not include performance components nor were they validated against external criteria of job proficiency. Indeed, the EEC was not staffed to handle performance testing or field validation.

In 1973, due largely to the influence of the performance-based training movement, the Army changed its approach to soldier evaluation by moving from norm-referenced paper-and-pencil tests to criterion-referenced performance tests. The new tests were called Skill Qualification Tests or SQT. Test results were still to be used for personnel management purposes but the primary focus of SQT was redirected toward the training and combat readiness of individual soldiers.

The overriding requirement of SQT was that they be job relevant. Test content was tied to critical job tasks that were identified through job and task analysis and described in the Soldier's Manual given to each soldier.

Performance was tested by one of three methods: (1) hands-on--a standard performance test in which the task conditions are simulated and the soldier demonstrates performance; (2) written knowledge--a multiple-choice test about critical elements of task performance; and (3) performance certification--a task-based evaluation conducted by the commander in the actual job setting. These alternatives were intended to give the test developer needed flexibility in accommodating tasks with different behavioral characteristics and different situational support requirements.

Methods of Performance Measurement

The evalution of performance measurement in the Army gives rise to an issue that is central to the conduct of Task 5 research and development activities: appropriate methods of measurement. Issues concerning the choice of methods generally center around trade-offs between the cost and validity of alternative approaches. The more precisely one specifies the performance to be observed and the conditions under which it is to be observed, the higher the cost.

Frederickson (1962) and Engel (1970) offer simple taxonomies of performance evaluation measures. Both tend to distinguish measures along two continua of remoteness or indirectness relative to actual job performance: the remoteness of the test behavior observed and the remoteness of the observer or scorer. Job performance tests are generally viewed as the most direct

method since they call for application of knowledge and demonstration of skill by eliciting behaviors that are equivalent, or nearly equivalent, to those required in the job setting. But the directness of this method--with its inherent relevance, content validity and fairness--comes at a price. Many personnel managers believe the benefits of performance testing do not justify the demands on facilities and personnel (Harris & Mackie, 1962) nor the wear and tear on equipment (Angell et al., 1964). Also, the level of professional skill available in the military to develop and administer performance tests has been questioned (Vineberg & Taylor, 1972). And yet another shortcoming of performance tests--obvious but not widely discussed--is that the greater administrative time they require usually restricts coverage of the job domain; one can measure fewer job tasks per unit of time than is possible with less direct measures.

The shortcomings of performance tests, especially that of cost, have led to the widespread use of job knowledge tests. Job knowledge tests consist of questions about task performance, usually delivered in a paper-and-pencil multiple-choice format. They are indirect measures to the extent that the behaviors measured do not constitute task performance but only mediate it.

Despite their evident economy, a question lingers concerning the degree to which knowledge tests can adequately gauge a person's job performance

capability--either in terms of the range of job behaviors that can be validly represented by knowledge items, or in the sense that knowledge testing in a paper-and-pencil mode presumes at least minimal literacy. Shirkey (1966), Urry, Shirkey and Waldkoetter (1965), and Yellen (1966) found correlations between job knowledge test scores and work-sample criteria to be too low to support the use of knowledge tests alone to assess individual proficiency in the MOS for medical specialist, supply specialist, cook, and truck vehicle mechanic. Similar results were also obtained in Engel and Rehder's (1970) study of general vehicle repairmen, and Foley's (1974) review of the research on maintenance performance. On the other hand, knowledge tests do appear to have adequate validity for jobs with minimal motor-skill demands (e.g., personnel specialist) provided that only knowledge actually required on the job is covered in the test (Urry, Shirkey & Nicewander, 1965; Vineberg & Taylor, 1972). Adequate validity also was observed in a more recent study by Osborn and Ford (1977) in which the knowledge tests were evaluated against a hands-on mastery criterion for low-skill manual tasks. Controlling for mental ability and level of task mastery, correlations on the order of .70 were found between various kinds of knowledge tests and hands-on task performance. These high correlations, it is important to note, were attributable to two factors: (1) the skilled aspect of the tasks tested consisted essentially of recalling functions, not of manual performance, making a knowledge medium appropriate; and (2) the knowledge items were meticulously tied to the critical steps in task performance through a careful task analysis.

Affective classes of behavior, such as motivation to perform a task, <u>can</u> be assessed by performance tests if one uses unobtrusive measures (Osborn &

Ford, 1977). But to embed a task in some simulated job context sufficiently broad to permit the task to be performed voluntarily requires time and expense not typically justifiable. Standardization and scoring problems also militate against attempting to test motivational behaviors _in situ_ (e.g., Harris et al., 1975).

Similarly, time pressures, inadequate supplies and equipment and lack of peer or supervisory support can all influence performance of soldiers who otherwise know how and want to do the job correctly. An indirect measure, usually in the form of a rating by a supervisor or peer, is therefore considered a more feasible method of tapping the affective or "will do" aspects of job behavior. Supervisor or commander ratings typically do not correlate highly with job knowledge or job sample test performance (e.g., Engel & Rehder, 1970; Vineberg & Taylor, 1972), but this does not rule out their use for measuring aspects of performance not represented in knowledge or hands-on tests. Such ratings can be particularly useful when developed in ways that anchor the rater's judgments to specific, relevant job behaviors (e.g., Borman, Dunnette & Johnson, 1974; Borman, Hough & Dunnette, 1976; Campbell, Dunnette, Arvey & Hellervik, 1973; Toquam & Borman, 1981).

Job and Task Analysis in the Army

Procedures for systematic job and task analyses were developed in the 1950s, largely as the result of research conducted by the U.S. Air Force (e.g., Miller, 1953; Van Cott, Berkun & Purifoy, 1955; Christal, 1969, 1974). These procedures have been widely used by the Army in support of the system engineering of individual training, as articulated in the

Instructional Systems Development (ISD) approach (e.g., TRADOC Pam 350-30, 1975). A number of documents has been produced that provide guidance in applying these procedures to Army jobs and tasks (e.g., CON Reg 350-100-1, 1972; CON Pam 350-11, 1973; TC 21-5-7, 1977; TRADOC Cir 351-28, 1978; and TRADOC Pam 351-4 (T), 1979).

The Army's use of job analysis procedures has tended to be training oriented. That is, the information provided has been used largely to help make decisions about the need for and content of training given in AIT and other specialized courses. While other activities have also reflected and benefited from the knowledge gained from task analytic work, (e.g., designing job aids, developing SQT, constructing selection batteries, preparing job-related handbooks and manuals) the primary thrust for task analytic information has come from the various proponent schools. And, in fact, they have the primary responsibility for carrying out the task analyses for their own MOS. The training emphasis of task analytic work has important implications for the work to be done in Task 5.

The task data collection procedure most favored by the Army and other services is the job inventory, a standardized and self-administered checklist. It is the method of choice because interviews, observation-interviews, technical conferences and open-ended questionnaires have too many limitations to be useful in any large-scale data collection program (Rupe, 1956). It is the approach currently lying at the heart of the Army Occupational Survey Program (AOSP).

The checklist contains items describing a variety of duties and tasks related to a given MOS. These items are drawn from information already known about the job, primarily from existing documentation and from SME. (Guidelines for constructing a job inventory have been described by March and Archer [1967] and are included in TRADOC Pam 351-4.) Soldiers who are incumbents of the target MOS are instructed to check the duties and tasks that they perform, and to rate them on one or more dimensions such as frequency of performance and the related amount of time that they require to perform. Research has shown that incumbents are the best source of job inventory data; their supervisors do not have sufficiently precise knowledge of how duties and tasks differ in terms of time spent or other dimensions (Madden, Hazel & Christal, 1964).

A quantitative assessment of job activities can be obtained from a statistical analysis of the checklist responses using the Comprehensive Occupational Data Analysis Program (CODAP). CODAP can be used to rank-order duties and tasks in accordance with the percentage of soldiers who perform them and the relative time spent on each. This information, when combined with a number of other factors, is often used to select the critical tasks that will be the focus of training and evaluation activities. Survey Reports are prepared by the U.S. Army Soldier Support Center that summarize the results of the surveys for each MOS. These reports include valuable information on the structure and nature of the MOS by skill level. A major use that is made of these analyses is to determine which of the tasks comprising an MOS should be taught in a formal school setting (e.g., in AIT rather than on-the-job). Once these tasks have been selected, they are subjected to more detailed scrutiny to determine (1) how

they are best taught, and (2) the nature of the requirements they impose on the trainee.

Relevant Job and Task Analysis Methods

Job and task analysis (TA) activities are at the core of work to be done in Task 5. While the development of appropriate performance measures is the primary product of the task, the job and task analysis activities to be carried out are the primary input to that product. Several issues need to be kept in mind as the work connected with these activities is carried out. These issues involve accuracy, completeness, and appropriateness of job and task analysis activities.

In considering these issues, we need always to balance two factors that tend to work against each other -- the economy and efficiency of using existing Army job and task information vs. the need to supplement such information with new inputs from the project itself. Obviously, time and cost issues must be taken into account in coming to a proper balance between conducting new task analyses and supplementing existing task analyses.

Montemerlo and Eddowes (1978) reviewed more than 100 "how-to-do it" manuals for job and task analyses developed between 1950 and 1976, many of them military in origin. Each of the manuals was designed to proceduralize and systematize the collection and analysis of job and task information. The authors contend that such efforts have led to an oversimplification of the process so that the information obtained often does not accurately reflect

the job and tasks being analyzed. Miller (1973) expressed this concern very well:

> The space allowed (on the TA form) tyrannizes the space used, which in turn tyrannizes what information will be entered, which tyrannizes what information one will think about for entering into the format.

Such concerns have been expressed most specifically for those jobs (MOS) or parts of jobs (duty positions) that are not "proceduralizable." These kinds of activities are complex and interact synergistically with other tasks and sub-tasks within or outside the job. The danger is that the task analysis procedure itself tends to convert the subtleties of these jobs into a set of invariantly ordered steps, thus distorting the true nature of the job. Performance tests developed from such an analysis similarly would distort the criterion measures.

Another concern related to accuracy and completeness is the reliance placed on existing documentation and records. Prelewicz (1977) reviewed the task analytic work carried on at eight different proponent schools, and noted the following problems:

(1) Those who do the work are often not adequately trained.

(2) Those in supervisory positions sometimes change the results without consultation with the analysts.

(3) The analysts do not reflect the point of view of the job incumbent and resist change.

(4) Analysts think in "big chunks" rather than at the "how to do it level," thereby missing important information.

(5) Conditions and standards are not considered in detail or are given cursory treatment. Standards are sometimes "made up" because you need to say something. "Must

complete task in six minutes," may be a completely irrelevant "requirement."

(6) Too heavy a reliance is placed on the inputs of the proponent schools, including previous TA work and documentation, and local (school) SME.

(7) Not enough time is spent with current job incumbents, observing what they actually do in the field.

(8) Command emphasis on TA work is often not adequate to allow the school to do the job properly. Time, money, and personnel are not available in sufficient amounts.

By focusing TA activities on training concerns, analysts tend to look at how job/tasks are carried out at the novice or trainee level rather than at the level of the highly skilled or professional performer (Klein, 1978). The underlying assumption is that skilled performance is simply unskilled performance done better. The novice is said to be simply slower, makes more errors, and does not attend to the proper stimuli or cues for initiating or terminating steps in the task. This is the building block approach to skill/proficiency development. The counter argument notes that the proficient person does things <u>differently</u> than the novice -- that a new set of skills evolves out of the earlier ones (DeMaio, et al., 1976; Knoop & Welde, 1973; Klein, 1976). Thus, one needs to capture and describe the skilled performer's behavior in order to be able to (1) develop adequate measures of his/her performance and (2) to be able to develop predictors of that performance.

This problem as stated is consistent with the findings of Rose, Shettel, and Wheaton (1981), in studying the relationship between tasks as they are described in selected Soldier's Manuals (a product of TA), those tasks as measured by the SQT, and those same tasks as actually performed on the

job. In fact, 73% of the 1,223 soldiers from whom data were collected noted that tasks in the Soldier's Manuals differed from the way they are done on the job. Many of the soldiers noted that this difference often reflected the way in which the job should be done by a trainee while learning and the way it efficiently and effectively can be done by a skilled practitioner.

It seems clear that reliance on existing tasks and job information must be tempered with a number of cautions. We need to ensure that we have current, complete, accurate, and relevant information that serves the needs of both the developers of criterion performance measures and of predictor batteries. This means that existing task analytic information needs to be verified independently by SME from schools and field units and by comparing information contained in different Army documents.

Different job and task analysis methods are best suited for different things. An analysis of such alternatives was carried out in 1974 by Brumback, Romashko, Hahn, and Fleishman. Five job analysis methods were evaluated on the basis of 13 criteria. These methods included the job inventory approach used by the military (Christal, 1969), the U.S. Department of Labor Functional Job Analysis approach (Fine, 1955), the Position Analysis Questionnaire (PAQ) of McCormick (1972), the Fleishman Abilities Analysis approach (Fleishman, 1972), and the Critical Incident Technique (Flanagan, 1954). The conclusion drawn by the authors is that no method is uniformly the best. Each is weak in at least one respect compared to the others.

In any complex job setting a multi-method approach therefore is indicated in which requirements are determined both quantitatively (using existing Army job inventory procedures) and more qualitatively (by applying one or more of the more judgmental approaches). This suggests an eclectic approach that allows borrowing from different job analysis methods those parts that best serve specific project needs. Thus, we will, for example, use the critical incident technique to develop criterion rating scales but use the AOSP job inventory approach to identify important tasks, etc. In this way, we believe that we can compensate for the concerns relating to the accuracy, completeness and appropriateness of the task and job information obtained in Task 5, while at the same time remaining within established time and cost parameters.

It seems that different methods of performance measurement have different advantages and disadvantages. Despite their cost, hands-on performance tests, correctly developed and administered, cannot be equalled in job relevance, fairness, or acceptability to examinees (Schmidt et al., 1977); nor is there a known substitute for a performance test in measuring proficiency on tasks involving psychomotor skill. Knowledge tests, if used for the right kinds of job tasks and linked methodically to knowledge-based task elements, have wide applicability, acceptable validity and are exceptionally efficient. Performance ratings are the most remote measures, but permit the measurement of affective dimensions that cannot be feasibly tapped by other means.

SPECIFIC OBJECTIVES

The specific objective of Task 5 is to develop reliable, valid and economical measures of first and second tour job performance of enlisted personnel in a sample of nine MOS. These measures will serve both as:

(1) Data collection instruments for establishing the relationships among various kinds of predictors and criterion measures, and

(2) Prototypes for the development of performance measures for additional MOS and/or MOS clusters.

Two different kinds of performance measures will be developed. The first will be direct measures of task performance (e.g., the average time it takes a soldier to troubleshoot and repair a malfunctioning electrical component). For measures of this kind, the incumbents must be evaluated under carefully structured and standardized conditions. The second kind will consist of measures that are based on indirect evidence of performance (knowledge tests and ratings by supervisors or peers).

Both kinds of measures are needed. Instruments of the second, cheaper type are needed for operational use in monitoring performance, and for the Army's continuing efforts to improve selection and classification (which will not end with this project). Instruments of the first type are needed in order to develop the second. They also are needed to calibrate periodically the accuracy of selected predictor instruments. The careful calculations of utility that will be made in this project would be open to serious challenge if they were based solely on less direct measures of performance.

OVERALL SUMMARY OF PROCEDURE

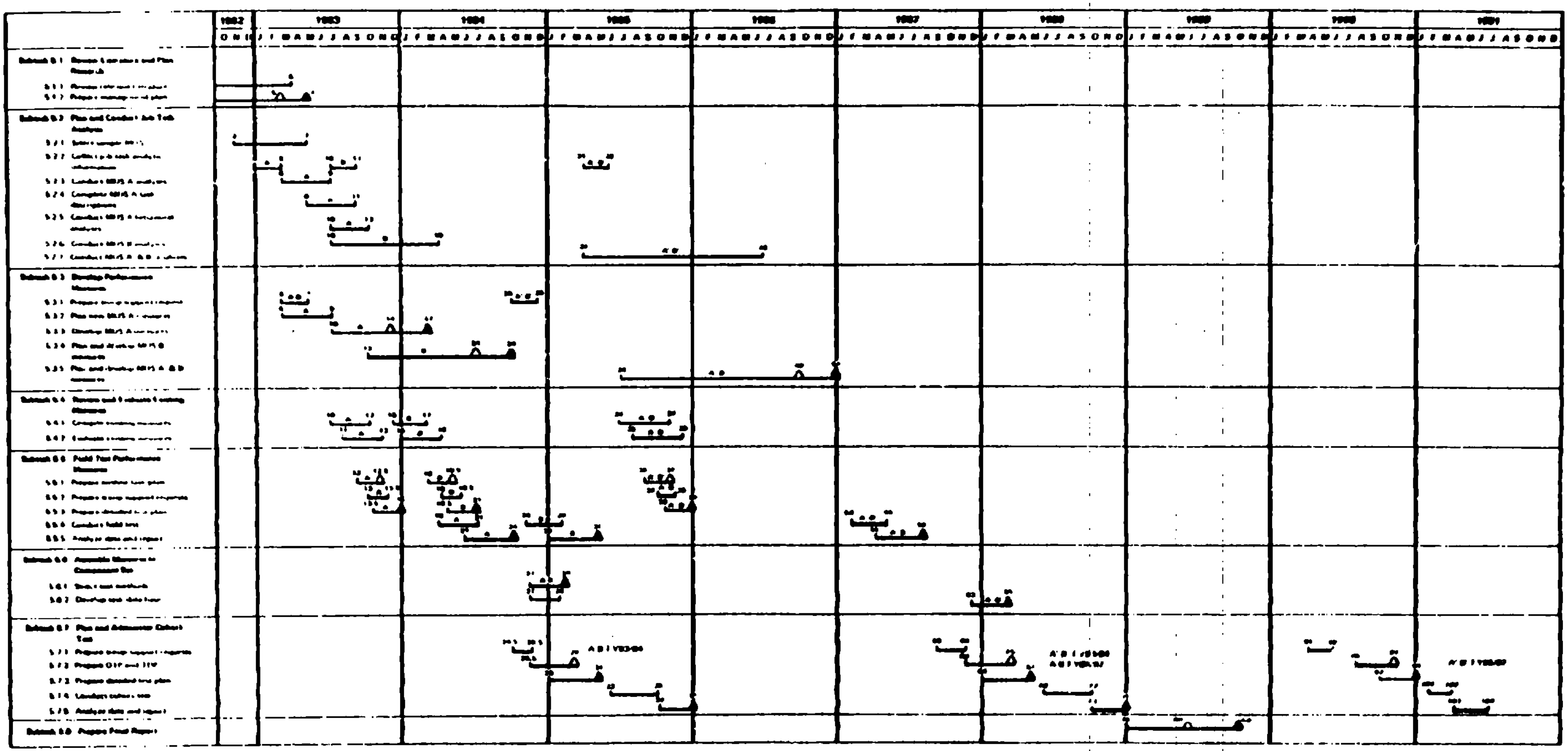
The subtasks, activities and milestones for accomplishing Task 5 are shown in Figure 5-1. The work begins in Subtask 1 with a review of the literature to identify job analysis and performance measurement methodologies for use by researchers in subsequent analytic and development phases of the project. Research and staffing plans will also be prepared in this subtask.

The job and task analyses in Subtask 2 serve four purposes. First, to provide Task 2 researchers with information about the criterion constructs underlying a variety of jobs within and across MOS, we begin with an analysis of the MOS performance domain. Second, to describe and analyze the task content of the MOS-specific performance domain, we will identify the duties and tasks performed by soldiers in the MOS selected for the research. Third, we will describe and analyze each task for which a performance test will be developed. Finally, we will describe and analyze the performance of soldiers on the job in terms of underlying dimensions that distinguish superior, successful, and unsuccessful incumbents. The group of MOS for evaluation will be specified in the context of this subtask.

The inputs to Subtask 3 are the recommended performance measurement techniques from the literature review in Subtask 1 and the designated MOS, tasks, task descriptions, and behavioral analyses from Subtask 2. Subtask 3 products are new performance measures of task-specific and general job

Table 5-1

Task 6 Schedule



dimensions of MOS proficiency. The measures and administrative support materials for their testing will be submitted for COR review.

The objective of Subtask 4 is to compile existing MOS performance measures and to evaluate their usefulness. Given the existing and the approved new measures we will begin field testing (Subtask 5) to assess their quality. The measures of MOS-specific performance will be evaluated in terms of psychometric considerations, content and construct validity, and practical utility and cost.

In Subtask 6, based on the results of the field tests, we will assemble existing and new performance measures into a component set. In Subtask 7, we will use the component set of measures to test two cohorts, the first in FY83/84 and the second in FY86/87. The proposed schedule enables us to develop, field test, and compile measures in time to apply them to the FY83/84 cohort before the incumbents finish their first tour. We plan a longitudinal design in which we will obtain performance measures on those soldiers in the cohort who reenlist and are available in their second term. Finally, we will replicate the FY83/84 cohort data collection effort by testing soldiers in the FY86/87 cohort in their first and second terms. Throughout, Task 5 data collection efforts will be coordinated with those of Tasks 2, 3, and 4. This will enhance the integration of research on pre-induction predictors, training/school measures, and Army-wide performance measures with the job-specific criterion data, and reduce demands on Army resources as well.

The research activities of job analysis, task analysis, performance measurement development, review of existing measures, field tests, and assembly of component sets of measures have been partioned into three partially overlapping phases. Measures are first developed for 4 MOS, Batch A at skill level 1 (SL1). Then, measures for the remaining 5 MOS, (Batch B) at skill level 1 are developed, followed by the development of skill level 2, (SL2) measures for the 9 MOS (Batch A'B'). Figure 5-1 depicts this iterative approach for each subtask. For example, we will begin the analysis of the second group of five MOS (MOS B) as we complete the development of measures for MOS A, but we will not begin to develop measures for MOS B until we obtain approval of the measures of MOS A. This step-wise schedule allows us to progress efficiently while obtaining approval of each set of measures. Detailed descriptions of plans for carrying out each subtask follow.

Before launching into discussion of those plans, however, we need to state a general operating principle that underlies this and each of the other research plans. The task will not be conducted in a vacuum where staff are impervious to the needs of or inputs from others. Below we anticipate and describe formal interactions with Tasks 1, 2, 3, and 4 in which we will share plans, proposal activities, and research results. Significantly, these exchanges will occur throughout the life of the project.

PROCEDURE

Subtask 5.1: Review Literature and Plan Research

The cornerstone for Task 5 work is a complete and easy-to-access literature file that addresses state-of-the-art methods of task analysis and performance measurement. Another prerequisite is a detailed plan for the research, to include its management and staffing. Thus, in this first subtask, two activities are planned: to review the relevant literature and to prepare research and manangement plans.

Activity 5.1.1 Review relevant literature. Drawing on the library resources of the three contractor organizations and using their reference accessing capabilities, documents will be compiled that pertain to: (a) job, task, and behavioral analysis, and (b) job performance measurement. Theoretical and empirical work on methods of analysis and measurement will be reviewed along with the results of pertinent Army applications in the form of completed job and task analyses and job-task performance tests. These documents and data will be reviewed and evaluated for relevance to the project. We especially will seek techniques and methods that can supplement existing Army information and current procedures. These documents and data of interest will be abstracted and catalogued for use by the Task 5 staff and the rest of the project staff. The job and task analysis review will focus on identifying methods that will enable us to: (a) define the Army job performance domain in terms of constructs that can guide the selection of predictors in Task 2; (b) partition job behavior

into tasks or dimensions of performance that best represent that job; and (c) detail the job tasks or behaviors in ways that provide for the methodical development of performance measures. The performance measurement review will focus on a comparative evaluation of different methods of testing and rating job performance. The reviews will be completed by the sixth project month.

Activity 5.1.2 Prepare research plan. A draft research plan has been prepared. The plan will describe the major subtasks and activities that will be performed, the interrelationships among the activities both within and across the subtasks and with those in other tasks, the schedule of task accomplishment, and the troop support requirements. The plan was revised on the basis of comments received from inhouse reviewers, the COR and from the Project A Advisory Groups.

Activity 5.1.3 Prepare management plan. To support the technical research plan, a corresponding plan for managing its execution was prepared. The plan allocates and budgets the staff, travel, material, service and overhead resources required to complete each Task 5 subtask. The plan, along with those for the other project tasks, will serve as input to the management information system through which project costs and progress will be monitored. A draft of this plan was completed by the fifth project month; the final plan is to be completed by the end of the seventh month.

Subtask 5.2: Plan and Conduct Job and Task Analyses

The goal of this subtask is to establish the critical elements of effective MOS specific job performance. This information is essential to the completion of Tasks 2, 3, and 5. The specific objectives of this subtask are:

(1) To identify the sample MOS for the overall project research, and the subset of 9 MOS for Task 5 research.

(2) To supply Task 2 researchers with information about criterion constructs underlying performance of a variety of jobs, both within and across MOS.

(3) To describe and analyze the performance of soldiers on the job in terms of those underlying dimensions that distinguish among superior, successful, and unsuccessful job incumbents.

(4) To describe and job-analyze the MOS-specific performance domain in terms of its task content.

(5) To describe and analyze each task that is to be represented by one or more job performance criterion measures in terms of how each task is performed in the job setting.

Subtask 5.2 will consist of seven major activities:

(1) Selecting a sample of MOS.

(2) Collecting MOS-specific and job/task analytic information (MOS A).

(3) Conducting job and task analyses on MOS A.

(4) Completing task description on MOS A.

(5) Conducting behavioral analysis on MOS A.

(6) Conducting job, task, and behavioral analyses on MOS B.

(7) Conducting job, task, and behavioral analyses on MOS A' and B' (second tour).

We describe each of these steps in turn. Each step implicitly contains an internal review phase. Each step will be completed when its final products have been approved by the appropriate senior staff.

<u>Activity 5.2.1 Cluster MOS and select the sample(s) of MOS</u>. A provisional sample of 19 MOS has been identified (see Table 1). The overriding criteria for the composition of the MOS sample are:

(1) That the number of job incumbents is large enough to produce reliable results from the data analysis; and

(2) That the variety of job skills collectively found in the sample be reasonably representative of the Army's job skill domain.

To meet these criteria, MOS were selected on the basis of (a) the number and mix of people to be trained in the job, and (b) the Career Management Field to which the job belonged, and (c) the representativeness of the MOS set of the types of jobs required to accomplish the Army's mission. The procedure generally entailed selecting a variety of CMF within strata of MOS density.

(1) A data table was generated listing for each Army MOS the number of troops acquired in FY81 and the number of those who are female, Black, or Hispanic.[1] The CMF to which an MOS belonged was also listed.

(2) A first pass was made through this table searching for MOS which had at least 1,000 troops overall and a minimum of 300 women, 300 Blacks, and 100 Hispanics. This pass produced 11 MOS in eight CMF. The first eight MOS were identified by selecting the largest from each CMF.

[1]FY81 accessions data were available. It was assumed that those data would represent reasonably well the relative distribution over MOS of accessions in FY83 and later.

(3) Next, the subgroup criteria were further relaxed by eliminating the requirement for Hispanic representation. This produced four additional MOS, but all were in CMF already present in the initial set of eight. On those grounds, all four were eliminated from further consideration.

(4) Again the criteria were changed, this time by eliminating the requirement for female representation but restoring the minimum requirement for 100 Hispanics. Against these constraints, eight new MOS surfaced representing four new CMF. Four MOS were added to the initial set of eight by retaining the largest in each new CMF.

(5) A final change in criteria was made in which the total accessions constraint was reduced from 1,000 to 500 and all requirements for minority representation were dropped. An additional 29 MOS in 14 CMF emerged. Seven of these 14 CMF were represented in the set of 12 MOS already selected. Of the remaining seven, one--CMF 98, Intelligence--was dropped because it is classified. That left eight MOS in six CMF. The largest MOS in each of the six remaining CMF was chosen, increasing our sample to 18.

A further indirect indication of the mix of job skills represented in the sample is in the range of ASVAB composites and component subtest pertinent to each MOS. All subtests and all but one (EL) of the nine composites were represented in the 18 MOS initially selected.

The extent to which ASVAB measures should be considered in evaluating the MOS sample for the range of job skills covered is debatable. On one hand, it seems tautological to choose or confirm the choice of criterion job skills on the basis of aptitude measures previously validated against such criteria. On the other hand, since one of the objectives of this project is to revalidate the present ASVAB against training achievement and job performance measures, it seems reasonable to choose a sample of MOS that gives ASVAB a fair chance for revalidation. Accordingly, we chose a 19th MOS (27E) which represented the EL aptitude composite.

The composition of the sample was also examined from the standpoint of mission criticality by comparing it with a list of 42 MOS identified by the Army as high priority for mobilization training.[2] The 42 MOS represent 17 CMF, 13 of which are contained within our set of 19. Of the four not in our sample, two are classified (CMF 96 and 98) and two are small (CMF 23 and 84). The six CMF in our sample not in the mobilization training priority list generally represent jobs for which there are civilian counterparts, a type of job purposely excluded from the mobilization list.

This initial set of 19 MOS represent 19 of the Army's 30 CMF.[3] It includes only 5 percent of Army jobs but 44 percent of the soldiers recruited in FY81. Similarly, of the 15 percent women in the 1981 cohort, 44 percent are represented in the sample; of the 27 percent Blacks, 44 percent are represented in the sample; and, of the 5 percent Hispanic, 43 percent are represented. While female and minority representation is high absolutely, relatively it remains about the same as in the population. The sample is 15 percent female, 27 percent Black, and 5 percent Hispanic.

Nine of the 19 MOS were tentatively earmarked for the job specific performance measurement phase of the project. These were selected, as a subset, with the same general criteria used in identifying the parent list of 19. Since the larger list is composed of five combat and 14 non-combat MOS, it seemed reasonable to see that these categories were represented in

[2] ODCSOPS (DAMO-ODM), DF, 2 Jul 82, Subject: IRR Training Priorities.

[3] Of the 11 CMF not represented, two are classified (CMF 96 and 98), two (CMF 33 and 74) have fewer than 500 FY81 accessions, and seven (CMF 23, 28, 29, 79, 81, 84, and 74) have fewer than 300 FY 81 accessions.

the subset of nine. It was further assumed that, to keep travel and field performance measurement costs within bounds, only the largest MOS be selected. So the three large combat MOS--11B (Infantryman), 13B (Cannon Crewman), and 19E/K (Tank Crewman)--were first selected. Of the 14 non-combat MOS, eight are large and have race and gender subgroups substantially represented. Since five different ASVAB composites are represented among the eight, one MOS was selected for each. Both 64C (Motor Transport Operator) and 94B (Food Service Specialist) share the OF aptitude composite and are roughly the same size, but the former was chosen because it is considered a priority MOS for mobilization. The two clerical (CL) MOS differ neither in size nor in their mobilization priority status, so 71L (Administration Specialist) was chosen over 76Y (Unit Supply Specialist) chiefly because it has more women. Both MOS with the ST composite were selected, since both have priority mobilization status. Thus, the nine MOS tentatively designated for Task 5 work are:

(1) 11B - Infantryman

(2) 13B - Cannon Crewman

(3) 19E - Tank Crewman

(4) 05C - Radio TT Operator

(5) 63B - Vehicle and Generator Mechanic

(6) 64C - Motor Transport Operator

(7) 71L - Administration Specialist

(8) 91B - Medical Care Specialist

(9) 95B - Military Police

An initial group of four, highlighted above, was selected and designated as Group A. While work will begin on Group A, the other MOS are subject to

further review. Lack of support for CMF as a job classification system is the main reason for this tentativeness. We have been unable to document the CMF structure as a systematically derived behavioral taxonomy of Army jobs.

As a check on CMF, we have undertaken a direct cluster analysis of MOS. Members of the contractor research staff and ARI Army officers--approximately 25 in all--have been given the task of sorting a sample of MOS into groups of their choosing based on perceived similarities and differences in job activities as described in AR 611-201. The sample of 111 MOS--which represents 47% of the population of 238 Skill Level 1, Active Army MOS with conventional ASVAB entrance requirements--includes the 84 large MOS (300 or more new job incumbents yearly) plus an additional 27 selected randomly but proportionately by CMF. Data from the sorting task were clustered and the initial results used to check the dispersion of our provisional sample of 19 MOS. On the basis of these results and guidance received from our Governance Advisory Group, two MOS that had tentatively been selected initially were replaced by 51B and 27E, which are in the same CMF and involve the same Aptitude Area Composites as the replaced MOS (62E and 31M).

The foregoing method of sampling provides MOS representative of the range of job skills in the MOS population while large enough for reliable estimation of individual test validities and differential validity across racial and gender groups. Yet, as stated in the Introduction, additional analyses of the MOS domain are required to support generalization of validities from the sample of 19 to the other 200 plus MOS. The next step in this direction will be to reaffirm the representativeness of the 19

through comprehensive cluster analysis or factor analysis of the MOS domain. Gaps in the sample of 19, revealed through this comprehensive analysis, can be filled by adding the necessary MOS to those researched in the 86/87 cohort. The procedure for this further MOS analytic work will, as described in the Introduction, be guided largely by the results of the pilot research underway presently. To allow time for development of measures for any new MOS, the comprehensive analysis of the MOS domain will have to be completed by the middle of 1985.

Activity 5.2.2 Collect MOS-specific and jcb/task analytic information. We will obtain job and task information specific to the selected MOS. This effort will take place during the two months that precede the planning of the job and task analyses of each wave of MOS. This will occur in:

(1) March-April 1983 for MOS A

(2) July-August 1983 for MOS B

(3) April-May 1985 for MOS A' and B'.

General information on enlisted MOS is available in the research team's libraries or has been obtained from Army sources such as MILPERCEN and the Soldier Support Center-National Capital Region. Some job and task descriptive data on the four initial MOS (MOS Group A) have already been obtained from the Army Occupational Survey Center. Additional specific information about each MOS selected for performance measurement will be obtained from at least the following Army agencies:

(1) Soldier Support Center - Army occupational survey reports and questionnaires; anticipated changes in MOS.

(2) Army Troop Support Center - Latest versions of Soldier's Manual, SQT (hands-on and written), duty position information; process for selecting which tasks to include in SM and SQT; anticipated changes in MOS task composition.

(3) Proponency Coordination Center - Issues, current or anticipated, that will affect task composition or duty positions of MOS, the distribution of troops in units or commands, or the topics and tasks trained in MOS schools.

(4) MOS Proponent Schools - Copies of current hands-on and written tests; task criticality lists; duty position information; anticipated changes in MOS task composition; relationship of tasks trained and tasks listed in SM and SQT; completed job and task analysis worksheets (TRADOC Form 550); Trainer's Guides.

(5) TRADOC Adjutant General - Educational Division - MOS task information from RCA Baseline Skills project.

Relevant documents and reports will be acquired and housed within the HumRRO project library.

Identify and analyze constructs and attributes. While Task 2 staff need a basis for tying their selection of predictors to criterion constructs, time and resources do not permit a comprehensive front-end analysis of the Army job domain. Data from four sources can be used, however, to provide a timely set of job-specific performance constructs.

One source is the outcome of the MOS cluster analysis described in Activity 5.2.1. These results may be used to select a representative but manageable number of MOS on which to focus an analyses of criterion constructs. This can be done, for instance, by selecting the two or three most representative (highest factor loading or index of belongingness) MOS from each factor or cluster for detailed analysis.

The form that this analysis takes will depend chiefly on the quality of data from a second source, the Army Occupational Survey Program (AOSP). The attractiveness of the AOSP's CODAP data is that a massive base of job task data for Army MOS can be compared, sorted, consolidated or otherwise examined by computer. Assuming that CODAP data are available on the sample of MOS mentioned, characteristics of the hundreds of job tasks in each can be analyzed to develop useful job performance constructs. There are probably many ways to summarize the data. One that we have begun to explore is to group tasks on the basis of action words. This can be done separately by MOS and then consolidated across MOS. Performance constructs stated in job activity terms -- "troubleshoots electronic/mechanical systems," "fills out forms," "engages targets," "assembles/disassembles mechanical equipment," "identifies targets," "cleans equipment," etc. -- may be determined in this way. Such clusters of job activities, when supplemented by dimensions from the behavioral analysis described in 5.2.5, should provide a useful set of job-specific performance constructs against which predictor constructs may be evaluated in Task 2 (Subtask 2.4). For delivery to Task 2, each construct will be named, defined briefly, clarified by examples and identified as to origin.

<u>Activity 5.2.3 Conduct job and task analysis on MOS A</u>. The relevant information for each MOS will be compiled, reviewed by staff, and prepared for analysis beginning in February and continuing through May 1983. Research staffs of AIR and HumRRO periodically will meet to plan the job and task analytic procedures that will be followed, to exchange preliminary findings of the analyses, and to review the final results. These meetings will insure consistent outcomes across different MOS. We will conduct the job and task analyses starting May 1983 and continuing through July 1983.

The terms "job analysis" and "task analysis," as used here, refer to a process of compiling existing information about each MOS (i.e., duty positions, tasks, task content and procedures), reconciling differences among various sources of information, and verifying the accuracy and validity of the revised job/task information.

To make certain that we have a realistic understanding of the MOS we are analyzing, we will make one- to two-day visits to nearby Army posts. We will observe troops performing the most frequently performed and essential tasks asociated with the MOS. Because of their proximity to AIR and HumRRO's offices, candidate sites for our visits include Ft. Knox, Ft. Belvoir, Ft. Meade, and the Aberdeen Proving Ground. These visits will not require the local command to provide any substantial personnel support. We would require only an escort at each post who could direct us to the appropriate work sites and, if necessary, explain the general nature of the work we would observe.

The initial phase of the job/task analysis will be to identify the duty position(s) to work with in the MOS. Our concern here is two-fold. We want to identify the duty position(s) with the largest number of incumbents in order to insure adequate numbers of troops who will comprise our testing samples. We want to choose duty positions that mirror characteristics which led to the selection of the MOS in Activity 5.2.1.

The first step is to identify the official and practical duty positions. The official duty positions are contained in AR 611-201. The official duty

positions may be subdivided by practical factors, most likely equipment. For example, the 13B MOS has 16 official duty positions. In practice the lead position, cannoneer, is divided further by type of gun. Four sources provide information on practical duty positions: the Soldier's Manual (SM), CODAP surveys, Trainer's Guides and SME at the proponent school and in units.

Once we have a general map of the MOS by duty positions at skill levels 1 and 2, we must judge the similarity of the positions. We would be most comfortable from a theoretical perspective if we only tested incumbents of the lead duty position. But that duty position may not have enough incumbents. Even if there are enough incumbents, we could get our sample from fewer units if the scope of the job to be evaluated were somewhat broader.

We will base our evaluation of the similarity of duty position primarily on the tasks and duties performed in each position. We will request the Army Occupational Survey Center to provide CODAP survey data reports showing tasks performed by duty position. These reports show tasks that are common across more than one duty position, as well as those specific to a single duty position. The CODAP task list will be augmented by task and duty position information from the SM, the Trainer's Guide and the proponent school. The resulting task-by-duty position list will be submitted for review by subject matter experts. The purpose of the review is to double-check the task list for possible recent changes in doctrine or practice. For this purpose, a few knowledgeable judgments are preferred to many marginally informed opinions. Thus, one or two NCO from the proponent school

who are familiar with current MOS doctrine and one or two from a FORSCOM unit who work daily with the MOS will be asked to review the list. As shown in Tables 5.2.1 - 5.2.4, the review is expected to take one day per NCO at each location.

We will select the target duty positions primarily on the basis of the number of incumbents in the lead duty position and those very similar to the lead position. AOSP provides the data for the official duty positions. If there are also practical duty positions, we will estimate the proportions of soldiers in each position. If a homogenous group of duty positions provides a suitably large sample of soldiers, performance measurement will be limited to those soldiers performing the tasks relevant to their duty position. If there is not a homogenous grouping, we have two choices:

(1) Track the data collection so that each soldier is tested not only on tasks common across duty positions but also on a sample of the tasks that make his duty position distinctive.

(2) Test only on the tasks that are common across duty positions. This alternative is less preferable because the supervisor ratings will be based largely on global performance of the distinctive tasks.

We will next identify the candidate tasks for which performance measures will be developed. Each task will be screened against these criteria:

(1) sufficient proportion of incumbents perform the task;

(2) The task is not likely to change or disappear in the immediate future;

(3) The task requires individual rather than team proficiency; and

(4) The task is deemed critical or important.

If criticality data from the proponent schools are not available to augment the basic CODAP data, we will secure them from MOS incumbents and their supervisors. In general, prime candidate tasks will be those that are difficult and important and performed by a large proportion of incumbents. The data will be obtained using a modified Nominal Group Technique conducted by a member of the project staff. SME will be asked to discuss and then independently rate a set of tasks for their criticality/centrality to the accomplishment of the MOS/duty position job. Fifteen SME per group (MOS) will be required, since experience with the technique indicates that too limited a perspective is represented by a group much smaller, and discussion becomes unwieldy with one much larger.

Project staff will categorize the tasks that survive the four-stage filter according to their functional content. We will then select tasks randomly to represent the proportion falling within each functional category. Our best guess is that the output will be a list of about 30 tasks.

Job analysis of MOS A will be completed by the 10th project month.

Activity 5.2.4 Complete task description of MOS A. The intent of this step is to describe in detail how each MOS task selected in the previous steps is performed. The task descriptions will consist of the task elements, task conditions and standards, and will be developed from information in relevant MOS school lesson plans, task descriptions generated by RCA in the Baseline Skills Project, SQT notices and tests. SM, Field Manuals, and Technical Manuals.

We will assess the sufficiency of each task description against these questions:

(1) Does the task statement describe observable and measurable behavior?

(2) Is the task the same if conditions vary?

(3) Are performance standards stated?

(4) Are performance standards appropriate to the duty position or skill level of the soldier?

(5) Are initiating stimuli identified?

(6) Are concluding stimuli identified?

(7) Is the use of references, job aids, memory aids, part of the task?

(8) Are all task steps or essential task elements listed?

(9) Is the level of descriptions consistent and conclusive?

The completed detailed task descriptions will be reviewed by SME. The SME will be two mid-level NCO instructors with recent troop experience. As mentioned before, evaluation of task data by an informed few is preferred to the opinion of many who are marginally informed. Two qualified SME who can check on one another's oversights and biases are a manageable number to work with. Thus, one project staff member who is familiar with our task descriptions will meet with the SME to review the descriptions task by task. Where conflict exists, we will require a compromise between the SME. For this review we will need two SME for each MOS for two days. (Tables 5.2.1 - 5.2.4).

The output of this step will be an approved detailed description of each task. These task descriptions will be the primary information used in developing performance measures (Subtask 3). They will be completed by the 12th project month.

Activity 5.2.5 Conduct behavioral analyses. In addition to the foregoing, we propose to conduct behavioral analyses (Borman, Dunnette & Johnson, 1974; Borman, Hough & Dunnette, 1976) of the MOS selected for investigation. The objectives of these analyses are: (a) to define in comprehensive, behavioral terms the performance requirements of these MOS; and (b) to develop rating scales that may be used to gather special for-research-only ratings to serve, in turn, as criterion performance scores in the predictor validation research. The procedure to accomplish the first objective is described below. Development of the rating scales is discussed in Subtask 3 with the other performance measures.

These job-specific rating scales are distinguished from those in Task 4 in that the latter are directed at Army performance in general. This distinction is conceptually clear; whether it holds up in practice remains to be seen. A separate set of scales may emerge for each MOS, or they may tend to converge toward a single set applicable to all MOS. But in either case, the Task 5 rating scale, in contrast to those in Task 4, are to be derived from behavioral incidents specific to MOS job performance.

Generate performance examples. As a first step in the behavioral analysis of a given MOS, we will identify soldiers and their supervisors (NCO) to participate in a series of one-day workshops to generate performance examples. Experience tells us that (a) we need about 1,000 performance example to be sure the job performance domain has been comprehensively defined, and (b) we can expect to get an average of about 10 usable examples from each soldier. Thus, a total of 100 participants per MOS will probably be needed. Participants should have at least two years'

experience in their MOS and should be those most fluent in oral and written expression. To keep the groups to manageable size, we propose conducting six such one-day workshops with about 16 participants in each. (These procedures and numbers are justified in Task 4 -- pp. 4-23 to 4-27.)

At each workshop, project staff will describe briefly the purpose of the research and then train participants to write behavioral examples. Next, participants will be asked to write examples of effective and ineffective job performance based on their experience with individuals in their own MOS. These examples take the form of short "stories" or vignettes about individuals performing on the job. Soldiers writing the behavioral examples will be encouraged to attend to the entire performance domain for their MOS when thinking about examples to write.

Edit performance example. The next step in the behavioral analysis is to edit the performance examples into a common format and to content analyze them to form preliminary performance dimensions. Once the dimensions are developed and defined, we will have them reviewed by the COR and a small number of persons knowledgeable about the MOS. This review will ensure that the dimensions make sense, are worded properly, and exhaust the target performance domain.

Review performance examples. At this point, we will administer by mail the edited performance examples and dimensions to the workshop participants. This step is designed to ensure that the performance scales are meaningful to persons knowledgeable about the target job. Specifically, job incumbents and/or their supervisors will review edited behavioral examples and

make two judgments about each. First, they will sort each example into one of the dimensions according to its content. Second, they will rate the effectiveness level it reflects (e.g., 1 = very ineffective to 7 = very effective). This procedure will point up ambiguities in the dimensional system or in individual behavioral examples if any exist. The result of this step will be a set of performance dimensions, well defined in terms of observable behavior. These data will then be analyzed to develop the final rating scales, as discussed in Subtask 3. The procedures for conducting behavioral analyses and developing anchored rating scales are described in more detail in the Task 4 Research Plan.

The behavioral analysis of MOS A will be completed by the end of 13th project month, the analysis for MOS B by the 18th month.

Activity 5.2.6 Conduct job and task analyses on MOS B. The same essential activities as described for MOS A will be repeated for MOS B. We expect MOS B will comprise five MOS. The job and task analytic work will be divided between HumRRO and AIR (PDRI will do the behavioral analyses under Activity 5.2.5). The estimated time frames are (dates are FY):

(1) Collect MOS B-specific information - 1 Jul - 30 Aug 1983

(2) Conduct job and task analyses on MOS B - 1 Sep 1983 - 15 Jan 1984

(3) Complete task description of MOS B - 16 Jan - 21 Feb 1984

Activity 5.2.7 Conduct job and task analyses on MOS A' and MOS B'.

Although the analysis of second tour MOS groups A' and B' will be conducted after the work on MOS groups A and B has been substantially completed, there will be considerable overlapping of the efforts. First, the conceptual thinking and the practical lessons of conducting the analysis on MOS groups A and B will necessarily affect the way that A' and B' analyses are carried out. Second, the consideration of differences between skill levels 1 and 2 within the selected MOS will be an implicit part of the initial analysis.

The job and task analyses for MOS A' and B' are scheduled for completion 30 June 1986.

Support requirements for Subtask 5.2. Soldier support requirements for job-task and behavioral analysis of MOS A, MOS B, MOS A' and MOS B' are shown in Tables 5.2.1 through 5.2.4.

Table 5.2.1

Soldier Support Requirements for MOS A Job-Task and Behavioral[a] Analyses

Purpose	MOS	Soldiers SL	Soldiers Number	Days Per Person
Review task distribution across duty positions	13B	3-5	3	1
	64C	3-5	3	1
	71L	3-5	3	1
	95B	3-5	3	1
Assess criticality of tasks (if necessary)	13B	3-5	15	1
	64C	3-5	15	1
	71L	3-5	15	1
	95B	3-5	15	1
Review task descriptions	13B	3-5	2	2
	64C	3-5	2	2
	71L	3-5	2	2
	95B	3-5	2	2
Provide critical incidents[b] and judgments for scale development	13B	2-5	100	2
	64C	2-5	100	2
	71L	2-5	100	2
	95B	2-5	100	2

[a]Skill Level 1 and Skill Level 2 behavioral analysis data obtained at the same time.

[b]SL1 soldiers with two years in service may be substituted for some of the SL2 soldiers.

Table 5.2.2

Soldier Support Requirements for MOS B Job-Task and Behavioral[a] Analyses

Purpose	MOS	Soldiers SL	Number	Days Per Person
Review task distribution across duty positions	5 MOS	3-5	3 per MOS	1
Assess criticality of tasks (if necessary)	5 MOS	3-5	15 per MOS	1
Review task descriptions	5 MOS	3-5	2 per MOS	2
Provide critical incidents[b] and judgments for scale development	5 MOS	2-5	100 per MOS	2

[a]Skill Level 1 and Skill Level 2 Behavioral Analysis data obtained at the same time.

[b]SL1 soldiers with two years in service may be substituted for some of the SL2 soldiers.

Table 5.2.3

Soldier Support Requirements for MOS A' Job-Task Analyses

Purpose	MOS	Soldiers SL	Soldiers Number	Days Per Person
Review task distribution across duty positions	13B	3-5	3	1
	64C	3-5	3	1
	71L	3-5	3	1
	95B	3-5	3	1
Assess criticality of tasks (if necessary)	13B	3-5	15	1
	64C	3-5	15	1
	71L	3-5	15	1
	95B	3-5	15	1
Review task descriptions	13B	3-5	2	2
	64C	3-5	2	2
	71L	3-5	2	2
	95B	3-5	2	2

Table 5.2.4

Soldier Support Requirements for MOS B' Job-Task Analyses

Purpose	MOS	Soldiers SL	Soldiers Number	Days Per Person
Review task distribution across duty positions	5 MOS	3-5	3 per MOS	1
Assess criticality of tasks (if necessary)	5 MOS	3-5	15 per MOS	1
Review task descriptions	5 MOS	3-5	2 per MOS	2

Subtask 5.3 Develop Performance Measures

Work on this subtask will proceed from results of Subtasks 5.1, 5.2, and 5.4. Specifically, the inputs will be comprised of recommended performance measurement techniques from the literature review in Subtask 5.1, the review of existing measures in Subtask 5.4, and the designated MOS, tasks, task descriptions and behavioral analyses from Subtask 5.2. Newly developed performance measures (covering both task-specific and more general job dimensions) constitute the output of this subtask.

Activity 5.3.1 Prepare troop support requests. Requests for support in developing the SL1 measures will be in the first Troop Support Request (TSR), submitted by the end of the 8th project month. Similar support for developing the Batch B and the second enlistment term measures will be included in later TSRs.

Activity 5.3.2 Prepare research and development plans for MOS A performance measures. Planning will be done and reported in three phases: rationale for new performance measures, procedures for development, and methods for evaluation.

The central question guiding performance measure development is: Given limited resources and access to a soldier for some fixed length of time, what aspects of job behavior should be measured, by what methods, in how many replications, in order to obtain the maximum amount of reliable data on the quality and efficiency of methods for measuring job-specific criterion performance? A series of guidelines will be established to match measures with tasks.

A detailed description of the procedures to be followed in developing each proposed performance measure will be included, along with examples of each type. In addition, the plans will suggest techniques for evaluating reliability, validity, cost and usefulness. The plans will be submitted to the ARI COR for evaluation at the end of the ninth project month. Review comments and recommendations will be followed by revisions, with final plans ready for implementation by the end of the eleventh month.

Activity 5.3.3 Develop MOS A performance measures. Three types of measures are planned: hands-on performance tests, performance-oriented knowledge tests, and behaviorally-based ratings. A fourth type of measure, computer-mediated knowledge tests, will be developed to the extent feasible.

Hands-on performance tests. Development work begins with the task descriptive data and proceeds through four steps:

(1) Determine scoring approach (process, product, or combination)

(2) Develop process items

(3) Develop product items

(4) Develop scorer's testing instructions.

The completed test package, which will consist of all tasks to be tested hands-on in a skill level, will be pilot tested with representative scorers and soldiers. The purpose of this is two-fold. The first is to assure that the test can be administered as designed in a field environment. The second is to determine scorer reliability.

Field acceptability will be checked by selecting a representative unit that contains the incumbents to be tested. Although the testing will be on a relatively small scale, the ability of the unit to support the tryout with the needed equipment, test site and scorers will give indications of the feasibility of the support specified.

Interscorer reliability will be established for each test by using a set of four representative scorers who score the performance of six representative soldiers. The percentage of agreement will be calculated as the number of actual agreements divided by the number of possible agreements. Any item on which there is disagreement among scorers will be discussed and considered for revision. The revision is likely to take one of four forms. The conditions may be changed to make a behavior more observable; a scoring aid may be added to facilitate more accuracy in measuring a product; the scoring instructions may be expanded to clarify the actions for the particular circumstance; or the item may be phrased more precisely. In addition, the scorer training materials will be revised to emphasize the procedure to score the item. All tests that include items that are revised because of low interrater agreement will be tried out again in another interrater reliability pilot test. If the item cannot be revised to produce an appropriate agreement level, the item will be deleted from the performance test.

In addition to the interrater reliability data, subjective data on acceptability and feasibility will be collected from scorers and examinees. Examinees will be asked whether they think their performance on the tests was a fair measure of their ability to do the task on the job.

Scorers will be asked if the standard and tolerance of the tests are consistent with their experiences, whether all necessary equipment was available, whether the scoring instructions were clear, and whether additional guidance was needed in response to any unanticipated incidents.

Performance-oriented knowledge tests. Paper-and-pencil tests of job knowledge, when compared to hands-on tests, not only provide wider coverage of the job domain at less cost in time and resources but also can prove acceptably valid for many job-tasks if the test questions are methodically anchored in task procedures. The sequence of decisions and actions to be followed in that anchoring hinge on the causes of failure to perform the task correctly. Each key behavior within the task will be analyzed rationally by staff and SME for potential causes of a failure:

(1) Is it because the soldier doesn't know WHERE to perform?

(2) Is it because the soldier doesn't know WHEN to perform a step?

(3) Is it because the soldier doesn't know WHAT the end result looks like?

(4) Is it because the soldier doesn't know HOW to execute the behavior?

For each likely cause of error, project staff and SME first will identify the correct location, or sequence, or product, or procedure; then describe it in words or pictures; then frame a question; and, finally, select real-world response alternatives (distractors) to complete the test item. The important point is that by considering these four questions about each aspect of task performance, we can pinpoint both what is important to ask in a knowledge test of task performance, and how to ask it. This procedure

helps prevent test questions that so often are used merely because they are easy to ask.

It should be noted that these knowledge tests differ in purpose and kind from those to be developed in Task 3. The latter are intended chiefly as training achievement measures to be administered before and after training. They are designed, moreover, to be comprehensive in the sense of addressing all tasks in the Soldier's Manual, but will do so by testing only a sample of task elements. The Task 5 knowledge tests, in contrast, are designed as potential substitutes for hands-on criterion measures. They will be developed only for a sample of job-tasks but will cover all essential performance elements for those tasks. Correlations between Task 3 and Task 5 knowledge tests, where common task elements are measured, will provide interesting data on the two approaches to job knowledge testing as well as trends in performance from school to the job.

Behaviorally-based rating scales. These scales, developed from the behavioral analyses described in 5.2.5, are aimed specifically at those aspects of the job that are particularly resistant to measurement by hands-on or knowledge tests, and they are designed to be free of the rating errors normally observed in conventional rating scales. The brief descriptions of performance (vignettes) obtained from soldiers will be edited, and then rated by soldiers as to the effectiveness of the behavior described. The data are used to prepare scales that pertain to dimensions of MOS-specific performance described by soldiers as important. Points along the scales are illustrated, with the vignettes, to help raters compare the behavior of the ratee with these benchmark behaviors. Thus,

soldiers themselves have provided the data to identify the dimensions, as well as to describe, by example, the various levels of performance. The scales will be tried out in the field test along with the other measures.

Computer-mediated knowledge tests. As a possible fourth method of testing, we propose to explore adapting the performance-oriented knowledge tests to a computer medium. If possible, these tests will be developed for a subset of the tasks covered by the job knowledge tests. Computer-mediated tests occur in other testing applications, but are seldom used in job proficiency assessment. Potential advantages of the approach are numerous and significant. The management of examinee response data is more efficient and reliable. Examinee responses to questions or test stimuli are recorded and processed instantly, enabling the dynamic management of test sequence and rapid, reliable reduction and reporting of test results.

Computer-mediated testing will be explored chiefly from the standpoint of its feasibility in terms of relative cost, range of task behaviors accommodated, and usability of the medium by soldiers. If considered promising, further development of computer-mediated versions of the performance tests will proceed in four phases:

(1) prepare test items
(2) identify system components
(3) develop software
(4) pre-test system

The majority of test items will consist of the performance-oriented knowledge items adapted directly to the computer-mediated format. Each job-task selected for testing will be reexamined in an effort to identify any tasks or task elements with reaction time constraints or visual motion cues that can be simulated effectively in a computer-mediated format.

Skill Level 1 measures for the first four MOS (MOS A) will be drafted amd submitted for approval by the end of the 14th project month.

Activity 5.3.4 Plan and develop MOS B measures. Skill Level 1 measures for the remaining five MOS (MOS B) will be developed following the same procedure outlined for MOS A. These measures will be drafted and submitted for approval by the 21st project month.

Activity 5.3.5 Plan and develop MOS A' and B' measures. Measures of second tour performance (Skill Level 2) will be developed for all nine MOS (MOS A' and B') and submitted for approval by the end of the 48th project month.

Support requirements for Subtask 5.3. A test developer and SME can develop a draft test (either hands-on or knowledge) for a typical task in about four days with the SME working half time on one test and half time on another. This includes reviewing the task analysis, developing the hands-on scoresheet and scorer instructions (or a sufficient number of knowledge items), and conducting tryouts. Our plan is to develop knowledge tests for all 30 tasks in each MOS/SL, and hands-on tests for half of these, making 45 tests in all. (These numbers are estimates arrived at by considering

potential trade-offs among number of MOS, number of tasks per MOS, number of measures per task, soldier support requirements, project design objectives, and project resources.) Two SME days for 45 tests is a total of 90 SME days for each of the MOS/SL shown in Tables 5.3.1 and 5.3.3. The SME should be at least one Skill Level higher than that for which the test is being developed. Four additional soldiers, who are similar to the SME but who have not participated in the development of the tests, will be required to review the knowledge items and to serve as hands-on test scorers in a scorer reliability study for each MOS/SL. Participation will total three days per scorer in order to cover five replications of the 15 hands-on tests. A minimum of six soldiers will be required per MOS/SL for preliminary tryouts of the instruments. Each subgroup of six should come from the MOS/SL being tested, but span a range of experience and, if possible, proficiency. They will be needed for three days to take the hands-on tests plus about a third of the knowledge tests. A second group of four scorers and six soldiers will be needed for two days to try out the revised instruments.

The number and kinds of soldiers needed to support development of MOS A performance measures are shown below in Table 5.3.1. The major items of equipment that we may need access to in order to develop the MOS measures are listed in Table 5.3.2.

The number and kinds of soldiers needed to support development of performance measures for MOS B, A', and B' are shown in Table 5.3.3.

Table 5.3.1

Soldier Support Requirements for Developing
MOS A Performance Measures

Purpose	Estimated Date	MOS	Soldiers SL	Soldiers Number	Days Per Person
Support test development	May-Jul 83	13B	2	3	30
		64C	2	3	30
		71L	2	3	30
		95B	2	3	30
Check scorer reliability	Jul 83	13B	1	6	3+2
			2	4	3+2
		64C	1	6	3+2
			2	4	3+2
		71L	1	6	3+2
			2	4	3+2
		95B	1	6	3+2
			2	4	3+2

Table 5.3.2

Possible Equipment Support Requirements for Developing
MOS A Performance Measures

13B
- M101A1 or M109 Cannon
- Direct fire telescope
- Panoramic telescope
- Collimator

64C
- 1/4-ton truck, utility, M151 series
- Truck trailer, 5-ton, M818 Series
- Semitrailer, stake and platform, 12-ton, M127 series

Common Soldier
- M60 machinegun
- M16A1 rifle
- M203 gredade launcher
- M18A1 Claymore mine inert

71L
- Typewiter

95B
- 1/4-ton truck, utility, M151 series
- FM radio set

Table 5.3.3

Soldier Support Requirements for Developing MOS B, A' and B' Performance Measures

Purpose	MOS	Soldiers SL	Number	Days Per Person
MOS B				
Support test development	5 MOS	2	3 per MOS	30
Check scorer reliability	5 MOS	1	6 per MOS	3+2
		2	4 per MOS	3+2
MOS A'				
Support test development	13B	3	3	30
	64C	3	3	30
	71L	3	3	30
	95B	3	3	30
Check scorer reliability	13B	2	6	3+2
		3	4	3+2
	64C	2	6	3+2
		3	4	3+2
	71L	2	6	3+2
		3	4	3+2
	95B	2	6	3+2
		3	4	3+2
MOS B'				
Support test development	5 MOS	3	3	30
Check scorer reliability	5 MOS	2	6 per MOS	3+2
	5 MOS	3	4 per MOS	3+2

Subtask 5.4 Review and Evaluate Existing MOS-Specific Measures

The goal of this subtask is to compile existing MOS A performance measures and to evaluate them with respect to their utility as indicators of job-specific performance. SQT for the MOS of interest are the most obvious example. These performance measures exist, and, if they meet certain criteria of acceptability, would obviate the necessity of developing a new test. Also, a "good" test developed in the school setting, while not appropriate as a job-specific measure, might be efficiently adapted for use as such.

The work on this subtask can begin when MOS A has been identified and the tasks selected for testing.

Activity 5.4.1 Compile existing measures. Measures that expand our coverage of the criterion space for an MOS without adding to testing time are not expected to be numerous. For a measure to be useful in this regard, it should be both comprehensive, in that it covers a significant sector of the criterion space, and already in operational use so that scores are available for the soldiers under investigation in this project. The two most obvious measures that meet these standards are the SQT and the Enlisted Evaluation Report (EER). We will query MILPERCEN and TRADOC regarding other such operational performance measures for the job specialties in MOS A.

Finding available measures that can be adapted to our purposes is much more likely. Here we are looking for existing tests or rating instruments, pertaining to tasks or behaviors identified for measurement in Subtask 5.2, which can be used to save development time or otherwise enhance the set of measures developed in Subtask 5.3. We will screen two major sources:

1. Tradoc EPMS Network. Performance measures developed within the TRADOC Enlisted Personnel Management System, chiefly by the Directorates of Training Developments (DTD) within the schools, will be compiled for review. These are measures developed typically for use in center, school and unit training evaluations, but which hold promise for adaptation to the broader purpose of performance appraisal.

2. ARI-Contractor Research Projects. Since many Army personnel research projects involve development of performance tests, ratings of performance, and other criterion measures, this is a potentially rich source of performance measures. Reports of research conducted by and for the Army Research Institute which are relevant to the target MOS will be identified and examined for useful performance measures. For example, research of the type produced by Shields, Hanser, Williams, and Popelka (1981) may provide some measures related to Skill Level 2 performance (although most of their measures are Army-wide).

Activity 5.4.2 Evaluate existing measures. Once identified, each relevant performance measure will be evaluated according to its intended use. Any intact, comprehensive, MOS-wide measure like an SQT will be evaluated in two ways. First, we will determine whether the measure can be administered and the scores made available for the job incumbents of interest and in a time frame consistent with other measurement. Second, we will decide if the measure is qualitatively acceptable and useful. This second stage of evaluation in turn will entail compiling and analyzing two kinds of data. First, we will need information pertaining to a measure's development (whether the measure or set of measures was developed according to sound practices). For example, we would need to determine whether prescribed procedures for SQT development have been followed. A second type of data to be examined is that resulting from operational administration of the measure. Summary statistics on SQT, for example, are available from the SQT Management Division (SMD) of the Army Training Support Center (ATSC). These data consist of detailed subtest information and item statistics and also indicate the tests' overall difficulty and range of performance produced.

Evaluation of measures considered for adaptation or use with the new set will be done similarly, but standards for accepting a measure will be more stringent. Generally, these standards would require evidence of development procedures consistent with those set forth in Subtask 3, in addition to persuasive data on the measure's demonstrated validity and reliability. Such rigorous standards will eliminate all but a few existing job-task measures from outright adoption; in other cases we may be able to rework an available measure into a new one suitable for tryout and field testing.

Subtask 5.5 Plan and Implement a Field Test of MOS-Specific Performance Measures

The field test will provide data to assess existing and new measures as criteria of MOS-specific performance. Elements of analysis include psychometric considerations, content coverage, practical utility and costs.

Three field tests are planned. The first, scheduled for project months 19-21, is to test the SL1 measures for the first batch of four MOS. The second, scheduled for months 28-30, is to test SL1 measures for the remaining five MOS. The final field test, scheduled for months 53-55, is to evaluate SL2 measures for all nine MOS.

Activity 5.5.1 Prepare outlines of test plans. The first activity in this subtask will be to initiate research coordination and troop support requests. This coordination will be effected by providing outlines of each test plan. The outline will spell out necessary administrative information and specify test objectives. In outline form, the plan communicates the nature and objectives of the test to scientific and military personnel who are responsible for approving and providing the troop support.

The outline will be followed up by a specific test design statement describing the scientific research aspects of the test. It will specify the conditions under which the performance measures must be tested, the experimental design, the data requirements, the analyses planned, and proposed use of findings.

Subtask 5.3 will provide much of the information for the outline and test design specification in the form of planning documents and the measures themselves. Thse documents will be combined after review by the COR.

Activity 5.5.2 Prepare troop support requests. The required troop support request will be submitted to allow a minimum of six months for processing plus schedule constraints (training schedules, holidays, summer National Guard training support, National Training Center exercises, etc.).

The data collection coordinator will submit the troop support requests in accordance with the Master Project Plan and, at the discretion of the COR, will follow through with briefings, telephone calls, and supplementary materials to the Army managers responsible for the troops requested.

The troop support request for the first field test will be submitted in the 8th month; for the second field test, in the 20th month; and for the third field test, in the 44th month.

Activity 5.5.3 Prepare detailed test document. This document guides the day-to-day operations of the field test. It presents to the Army personnel who support the research the description of their role by time and place. It also contains the data collection instruments and the procedures for quality control of the data on-site.

There are several different audiences for this product. The data collection coordinator, test site manager, Army test control officers, and COR use the entire document to guide and coordinate the data collection effort. Other users include test control officers, representatives of the supporting local units, hands-on test managers, and reseach assistants at the specific sites. To facilitate these uses, we will prepare local editions tailored to each test site and provide a table of contents to identify the sections for different users.

The test document for MOS A will be submitted to the COR for review and approval by the end of the fifteenth month.

<u>Activity 5.5.4 Conduct field tests of MOS-specific measures</u>. Implementation of the field test comprises three phases of activity: advance preparation on-site, execution of the test, and closure. We presume that the COR will forward the test plans to cognizant Army agencies including those that will provide the troop support. We will provide supplementary materials and coordination to facilitate the implementation. Advance preparation on-site requires approximately three days per test site for:

(1) briefings to the Commanders of the units supplying the troops to clarify the test objectives, activities, and requirements,

(2) examination of the test site, equipment, supplies and special requirements for the data collection and set-up of the hands-on test stations,

(3) training of the test administrators and scorers, and

(4) a dry run of the test procedures.

Successful test implementation requires that an officer of the supporting unit be assigned as test officer (e.g., a representative of the G3 or S3 office) and that a staff of NCO implement the controls for the flow of troops through the data collection procedures. We will review the logistics plan and test schedule with the unit's administrative staff and we will conduct the training of all civilian and military scorers and other data personnel. In the training phase, a dry run of the procedures will follow the data collection schedule and use the personnel and locations designated for the test. At the first test site, the dry run will evaluate the procedures as well as train the personnel. The training will focus on the handling of problem situations, particularly those requiring remediation by the scientific staff.

Because of the scope of the data collection activities for Task 5, this task will have a data collection coordinator who is highly skilled in Army field data collection. The data collection coordinator will manage the various field and cohort test implementations. Each test site will have a test site manager who supervises all of the research at an Army post during a field or cohort test. The test site manager is responsible for controlling the quality and flow of the data until delivery to the longitudinal research data base manager.

In addition to the test site manager, a project staff member, supported by a research assistant, will serve as the hands-on test manager for an MOS. (The exact number of MOS tested per site will depend on the distribution of incumbents by MOS and installation; once we know the exact MOS to be addressed in a field test, we can select the sites so that travel and

personnel resources are consolidated.) The hands-on test managers will have sufficient experience with field data collection to manage the Army personnel who serve as hands-on test scorers and others who assist with administration of the research.

Military personnel will serve as hands-on test scorers. The hands-on test scorers need to be familiar with the MOS tasks being tested. We prefer to have a cadre of NCO personnel for each MOS in the research. However, if designation of such a cadre is not possible we are prepared to train military personnel at each test site to score the hands-on performance tests.

If we must train hands-on test scorers at each site, we propose to use the existing system of test control personnel who administer the SQT systems. This approach will minimize the preparation needed for some of the research procedures and will reduce the burden on the Army.

We plan to conduct performance tests at several stations, to administer the set of measures for an entire MOS, and to complete testing of 30 soldiers in each MOS in 2-1/2 days[4]. Data collection for five MOS (150 incumbents per MOS) can be accomplished in two weeks at one site, with an additional week for site set-up, training, and collection of data from some of the group-administered tests. Support requirements include 25 NCOs/officers (an average of 5 per MOS).

---

[4]During the first field test, data will also be collected using Army-wide scales and knowledge and prototype measures developed by Tasks 4 and 3. The combined administration time will be two and a half days (see pages 4-44 and 3-53).

We will gather data other than those directly pertaining to the job performance measures. For example, we will examine the time and resources required for the various types of tests, the burden on the Army and on the soldiers, potential invasions of privacy, test credibility and other aspects of test acceptability. We will examine the relationship of the higher cost, individually administered performance tests to the less costly test types.

Closure of the field test has two major objectives. First, we will assure the quality of the field test data prior to leaving the site by identifying missing data points and obtaining the data as indicated. Second, we will debrief supporting units, reemphasizing the value of their contributions and providing what feedback we can on performance that may be requested by soldiers and commanders. We will probably return to the same posts to conduct the cohort tests. Our return will be facilitated by the good will of the unit personnel.

<u>Activity 5.5.5 Analyze field test data and report results</u>. During the field tests we will obtain data on a variety of measures that tap different aspects of MOS-specific performance. The data will be obtained for samples of approximately 150 soldiers not in the target cohort. The types of data that will be available are indicated below:

(1) hands-on performance test scores

(2) performance-oriented knowledge test scores

(3) overall rating of job performance
(supervisor, peer, self)

(4) ratings on behaviorally-anchored scales (supervisor, peer, self)

(5) performance measure acceptability rating

(6) job experience data (recency and frequency of task performance)--obtain through administration of a short questionnaire

(7) measurement cost data

(8) SQT scores (if available)

(9) other (e.g., demographic information)

(10) computer-mediated knowledge test scores.

We will plan the data collection and analysis with the advice and assistance of Task 1 staff who will also participate in the analyses. Several major kinds of analyses will be conducted as discussed below.

Clean data and develop descriptive statistics. Statistical analyses will begin with the data verification procedures described in Task 1 and conducted by the analysts in charge of the LRDB. Although we plan to check data sheets for missing data before the instruments leave the test site, some instances of missing data are bound to occur. These will be rectified by means of the special PROC IMPUTE missing data routine. Once the data are "cleaned," standard descriptive statistics will be computed for the samples (by MOS) and subsamples (e.g., by ethnic group and gender) for all variables. These will include means, variances, ranges, frequencies, etc. Appropriate transformations will be applied to seriously skewed or otherwise non-normal distributions to render the data suitable for further analyses.

Determine reliability. Internal consistency approaches to assessing reliability are generally inappropriate for job proficiency tests, since job skills and abilities are not homogeneous but tend to vary from one area of job performance to another. This is equally true for hands-on and knowledge tests. Test-retest methods of estimating reliability, are more appropriate, but often difficult to implement because of the additional personnel and time demands in retesting. Two problems are associated with test-retest approaches to performance test reliability. One is the practical difficulty with extending the retest interval. It is difficult to get soldiers back to the test site a second time. This argues for retesting the day of the first test rather than after several days, despite the fact that an interval of days between test administration is preferred to one of hours. A second problem is that the examinee is changed by taking a performance test. This is more so than with other types of tests. Seeing the results of their actions in the course of performing a task can provide cues for changing behavior on retesting.

We plan to use two approaches to reliability estimation in field testing the hands-on tests. First, we intend to obtain retest data by attempting to get at least half of the tested soldiers back after an interval of several days. To retest all soldiers on separate days would nearly double the time and resources planned for field testing; yet to reduce the number of tasks tested so there would be time to retest later the same day would result in too few soldiers per task. So, we plan to request that all soldiers tested return several days later for retesting. Attrition will probably reduce the original number (150) substantially, resulting in perhaps 80-100 soldiers on whom retest data are available.

As a second approach to estimating the reliability of the performance tests, we plan to explore ways in which test performance can be partitioned for evaluation in an analysis-of-variance context. If assumptions of independence and randomization can be met, factors such as time, task type, test station, scorer, etc., may be identified or introduced as variates in order to examine the generalizability of test performance over such sources of variance (Cronbach, et al., 1972).

A split-half technique will be used to estimate the reliability of the knowledge tests. Items pertaining to each task may be divided into two halves, scores for the halves separately totaled and correlated over examinees--the "stepped-up" correlation providing the estimate of test reliability.

Estimating the reliability of job ratings by supervisors, peers, and incumbents themselves, is somewhat less critical. We are less concerned with the reliability of individual raters than with the reliability or constancy of the ratings of a given soldier. Since we will obtain two peer and two supervisor ratings for each soldier rated, agreement among raters can be used to estimate the reliability of soldier ratings. For the self-ratings, it would not be appropraite to routinely ask soldiers to repeat the self-rating task. To the extent that the rating scales can be paired (i.e. each pair of rating scales is viewed as covering the same or very similar performance tasks), we can use the correlation between the paired scales to place a lower bound on the reliability of the two scales. Otherwise, a lower bound on reliability will be estimated using the

multiple correlation of the self ratings with the peer and supervisor ratings and with the performance and knowledge measures themselves.

The reliability of other measures (job experience data and background demographics) will be checked for a sample of soldiers by obtaining equivalent information from supervisors.

Determine validity. We will conduct analyses of content, construct and concurrent validity. Content validation is largely a matter of making certain that test elements match task elements revealed by the task analysis. Insuring content validity is an inherent part of the test development process. Thus, prior to the field test we will, with the assistance of SME, have carefully compared each proposed item (including performance standards, sequence, test conditions) with the task analysis data, to assure that all items are part of the job requirements.

Construct validity of the job-specific criterion measures will be examined in two ways. First, item, task or dimension scores--depending on the instrument--will be intercorrelated, factor analyzed, and the resulting factor structures compared between criterion measures. Second, where measures of performance on the same tasks were obtained on the same soldiers by more than one method, a multitrait-multimethod analysis will be performed in an attempt to identify criterion constructs that are stable across methods of measurement. This will clearly be possible for those tasks ("traits") tested by hands-on and knowledge methods. It is also possible that some existing measure of task performance such as SQT scores will be available for field test participants as an additional method to be

introduced in the analysis. And, though more remote, it may even be possible to include some part of the behavioral ratings as still another method, if tested task performance can be mapped readily into one or more dimensions of the behavioral rating scales.

The primary technique used in the analysis of construct validity will be the estimation of multitrait-multimethod parameters through the use of LISREL V models. We will rely on assistance from Task 1 staff who have considerable experience in the application of such models and who will coordinate their use across tasks. The basic approach of these models is to view each observed measure as resulting from a combination of underlying construct (trait) and method variables plus some error variation (estimated by the reliabilities). The analysis then produces estimates of the relative importance of each underlying variable for each observed measure and of the overall fit of the model (Joreskog & Sorbom, 1981).

The results of the multitrait-multimethod analyses will also address the question of the extent to which the various job performance measures tap distinct versus equivalent criterion dimensions. As Task 3 and 4 performance measures will be available for the same soldiers, the field tests will provide the first empirical opportunity for an investigation of the dimensionality of the criterion space. A considerable discussion of the relevant issues in such an analysis may be found in the Task 1 research plan. These analyses will attempt to determine the number of different measures required to adequately cover the criterion space in the main cohort administrations. The models being identified or developed in Task 1 to address this issue in the main cohort analyses will, to the extent possible, be applied here.

Concurrent validities will be determined from intercorrelations between "indirect" measures (i.e., knowledge tests and ratings) and the more "ultimate" measures (i.e., hands-on performance tests). Evidence concerning the relationship of the indirect measures to the direct ones will be evaluated so that we can recommend, as appropriate:

(1) research use of the direct (more costly) measures,

(2) operational implementation of the indirect (less costly) measures, or

(3) a mix of direct and indirect measures that maximizes the cost-benefit of the measurement system.

Search for bias. The first test for differences will be a comparison of group mean scores, by ethnic group, by sex, and by ethnic group and sex combined. Analysis of variance will be the statistical test of choice. If significant differences are obtained, we must ascertain whether they appear to be due to bias (see the Task 1 extensive discussion of the bias issue). If there is a suspicion of bias, we will examine the measures to determine whether changes can be made that would reduce or eliminate it. Possible changes include simplifying instructions or options in written tests, if language appears to be the problem; providing special tools or mechanical assists on a performance test that examinees report using even if not required by the technical manual, etc. Finally, it may be necessary to eliminate some items from the scoring system if no other way can be found to equalize apparently biased scores.

Other analyses. All of the analyses described above, when applied to task-level performance data, assume that the performance scores are valid

indicators. That is, low proficiency is indicative of low MOS-specific job performance. Clearly, therefore, if an incumbent does not perform some of the target tasks frequently (or has not done so recently) the meaning of a low test score is ambiguous. To assess the degree to which such an artifact may underlie the obtained performance data (and affect reliability, validity, and bias), we intend to replicate all of the statistics and analyses discussed above, using task-level performance data that have been adjusted by means of appropriate covariate procedures, to control for recency and frequency of task performance.

Prepare report. We will report the results of the field test of new and existing MOS-specific measures three months after the completion of the field test, and will revise it based on comments from the COR. The report will have a management section that summarizes the types of information of use to Army personnel managers and will have a scientific section in a format suitable for ARI publication or for submission to a psychological journal.

Support requirements for Subtask 5.5. The number of soldiers per MOS/SL (N = 150) being requested for the field test is considered, in a statistical sense, minimally acceptable. With a sample of 150, correlations--which will be computed among test methods, subtests, items, rating scales, job experience data and other variables--of .14 or larger will test as significantly different from zero (1-tailed test at .05 level) using standard statistical tests for the significance of sample correlation coefficients. Similarly, this sample size will enable us to be 95 percent confident that an estimate of the proportion passing a measure is accurate within plus or

minus .08. An N of 150 will also enable us to detect unanticipated, infrequent events or problems that may occur in connection with field testing instruments. A problem that occurs for only one soldier in 50, which could seriously affect the larger cohort administration, will have a 95 percent chance of being detected (the probability of the event not occurring, .98, raised to the power of 150 is just under .05, the probability of the event never occuring in our sample).

Soldier support requirements for the field test of MOS A, MOS B, MOS A', and MOS B' are shown in Tables 5.5.1 through 5.5.4.

Table 5.5.1

Soldier Support Requirements for MOS A Field Test[a]

Purpose	MOS	Soldiers SL	Number	Days Per Person
Provide liaison with tested units	Officer		1 per post	12
Coordinate equipment and subjects	NA	3-5	2 per post	12
Score hands-on tests	13B	2	3 per post	12
	64C	2	3 per post	12
	71L	2	3 per post	12
	95B	2	3 per post	12
Subjects for measures[b]	13B	1	150	2.5
	64C	1	150	2.5
	71L	1	150	2.5
	95B	1	150	2.5
Supervisor ratings[c]	13B	3-5	40	1
	64C	3-5	40	1
	71L	3-5	40	1
	95B	3-5	40	1

[a]It is unlikely we can get 150 in each MOS at any post. An average of 5 support personnel per MOS per post will be required. The most efficient approach would be to field test each MOS completely at any given post.

[b]The 2.5 days includes time for the administration of Task 3 and 4 measures.

[c]Supervisor of tested soldiers. This requirement is described in the Task 4 plan (see page 4-44).

Table 5.5.2

Soldier Support Requirements for MOS B Field Test[a]

Purpose	Soldiers MOS	Soldiers SL	Soldiers Number	Days Per Person
Provide liaison with tested units	Officer		1 per post	12
Coordinate equipment and subjects	NA	3-5	2 per post	12
Score hands-on tests	5 MOS	2	3 per post per MOS	12
Subjects for measures[b]	5 MOS	1	150 per MOS	2.5
Supervisor ratings[c]	5 MOS	3-5	40 per MOS	1

[a]It is unlikely we can get 150 in each MOS at any post. An average of 5 support personnel per MOS per post will be required. The most efficient approach would be to field test each MOS completely at any given post.

[b]The 2.5 days includes time for the administration of Task 3 and 4 measures.

[c]Supervisor of tested soldiers. This requirement is described in the Task 4 plan (see page 4-44).

Table 5.5.3

Soldier Support Requirements for MOS A' Field Test

Purpose	Soldiers MOS	Soldiers SL	Number	Days Per Person
Provide liaison with tested units	Officer		1 per post	12
Coordinate equipment and subjects	NA	3-5	2 per post	12
Score hands-on tests	13B	3	2 per post	12
	64C	3	2 per post	12
	71L	3	2 per post	12
	95B	3	2 per post	12
Subjects for measures	13B	2	150	2
	64C	2	150	2
	71L	2	150	2
	95B	2	150	2
Supervisor ratings[a]	13B	4-5	20	.5
	64C	4-5	20	.5
	71L	4-5	20	.5
	95B	4-5	20	.5

[a]Supervisor of tested soldiers.

Table 5.5.4

Soldier Support Requirements for MOS B' Field Test

Purpose	Soldiers MOS	Soldiers SL	Number	Days Per Person
Provide liaison with tested units	Officer		1 per post	12
Coordinate equipment and subjects	NA	3-5	2 per post	12
Score hands-on tests	5 MOS	3	2 per post per MOS	12
Subjects for measures	5 MOS	2	150 per MOS	2
Supervisor ratings[a]	5 MOS	4-5	20 per MOS	.5

[a]Supervisor of tested soldiers.

Subtask 5.6: Assemble Old and New Performance Measures Into Composite Sets

This subtask has two purposes:

(1) To establish a data base that permits performance measures previously developed in the project to be specified as measures for the same or similar tasks performed in other MOS.

(2) To determine the measure(s) for testing each task in the cohort data collections.

Activity 5.6.1 Develop a data base for comparing tasks. One of the existing weaknesses of the current AOSP is that task analyses are conducted at different sites by different people with the consequence that inconsistent terminology is fairly common. We will identify commonalities by structuring a matrix that arrays tasks against measurement techniques. The cell entries will consist of new and/or existing performance measures. The total matrix will be capable of being stored in the computer. Hard-copy of relevant parts of the matrix would be available.

This activity will consist of three steps:

(1) Collect relevant information and data concerning new, existing, and recommended performance measures for each task.

(2) Develop a consistent descriptive system or terminology to describe these measures, and their characteristics in a form suitable for Army-wide use.

(3) Organize the collected information in a data base in accordance with the above system, including appropriate cross-referencing and categorization of information/data.

Activity 5.6.2 Select test methods for tasks. The decisions in this activity center on the question: Which tasks should be tested by what method or measure during the cohort test phases?

The problem is to select test methods so the total pool of tests has the highest concurrent validity and acceptability consistent with feasibility constraints. The data and experiences resulting from the field tryouts will guide the selection of test method for each task that is to be tested during the cohort phase. Three criteria are involved in the decision: concurrent validity, acceptability, and feasibility.

Concurrent validity. The primary concern is to maximize concurrent validity as indicated by the coefficients found during the field trials. In practice this is a negative criterion: Performance on the tasks with the lowest correlations between the written/computer and the hands-on tests for the tasks should be measured using hands-on tests. We expect that some differences in correlations will be traceable to physical or psychomotor skills required in the hands-on version that are not mediated by the type of knowledge that is covered in the written/computer versions.

Acceptability. The second, and a secondary, indicator for hands-on testing is the spread between soldiers' expressed preference for the hands-on mode as against the written/computer mode. If other factors are equal, the greater the spread between preference for the hands-on test and for the written/computer test, the more likely that the task will be tested hands-on. This criterion will help assure that the total test is perceived as fair.

Feasibility. The third criterion for selecting tasks for hands-on testing is feasibility. The major consideration is to require only equipment that can be made available for the cohort test. A second consideration is the amount of information to be gained in about one day of testing. Tasks with repetitive operations and extraordinarily time consuming steps are less likely to be tested hands-on than their "richer" counterparts.

Subtask 5.7: Plan For and Administer MOS-Specific Performance Measures To Main Cohorts

This subtask provides the criterion data for the project. The new and improved measures developed in Subtasks 5.3 and 5.4 and refined in Subtasks 5.5 and 5.6 will be administered to soldiers in the target cohorts. The results will be supplied to staff working on different tasks and will be analyzed in several different ways, depending upon the particular task requirements. Thus, these data will be used as predictor and/or criterion data from the perspectives of Tasks 2, 3, and 4, and as raw input to continued reliability and validity analyses for Task 5.

Activity 5.7.1 Prepare main cohort troop support requests. As the test plans are formulated we will submit refined requests at least six months before the troops are needed. The elements of the troop support request are the same as for the field test described in Activity 5.5.2.

Activity 5.7.2 Prepare draft data collection plans.

Test plan outlines and detailed test documents for the administration of the performance measures to the main cohorts of enlisted personnel will be produced similiar in format to those developed for the field tests, (Activity 5.5.2). One important difference, however, is that we will use the LRDB data on the main cohorts to facilitate our sampling of the cohort personnel by determining the characteristics of the personnel in the selected MOS, their location, and other relevant features. Use of the LRDB to determine personnel locations will be especially beneficial for obtaining representative samples and for portions of the research plan that require repeated measures of soldiers as part of a longitudinal design (e.g., in their second tour).

Activity 5.7.3 Prepare final data collection plans. The final data collection plans will incorporate comments received from the COR on the draft plans and will reflect feedback information concerning the availability of the troop support requested earlier in the applicable TSR.

Activity 5.7.4 Conduct main cohort data collection. Implementation of the cohort test will be facilitated by our advance knowledge of the location and characteristics of the cohort samples. However, attrition, relocation and reassignment of the soldiers in the cohort create problems of obtaining suitable sample sizes, especially for repeated measures. We have organized the cohort data collection under the supervision of a data collection coordinator for the entire project. That manager will have a stable cadre

of project personnel who serve as test site managers, hands-on test managers, and research assistants. This organization of the data collection activities and personnel will decrease the impact of the research on the participating units and assure the standardization needed for data quality control. We will institute quality control procedures in the cohort data collection that are similar to those discussed for the field test.

The advance preparation, data collection procedures, and test site closure in the cohort test will be similar to those in the field tests. The Task 5 measures applied will be those assembled in Subtask 5.6 as the criteria for the products of Tasks 2, 3 and 4.

We will administer the MOS A and B Skill Level 1 measures to FY83/84 first term incumbents during the first cohort test. (We have as a target the testing of 650 soldiers in each of the 19 MOS--19 for Task 4 measures, 9 for Task 5 measures.) By the time of the second cohort test, in the 69th project month, we will have prepared and field tested the Skill Level 2 performance measures for all 9 MOS for which specific measures are being developed. These will be administered along with Task 4 Army-wide measures to FY83/84 second tour incumbents (100 per MOS) while at the same time administering Skill Level 1 measures to FY86/87 first tour incumbents (500-550 per MOS).

The final data collection will be in the second term of the FY86/87 cohort (the 9th project year), when 100 Skill Level 2 soldiers in each of the MOS will be tested. The performance measurement and analyses will be

conducted by the the same cadre of data collection and analytic personnel who worked on the data in the preceding major administrations and analyses.

Throughout the cohort tests, the data collection activities will be coordinated to integrate the research on pre-induction predictors, training measures, and Army-wide performance measures with the final criterion measures and to reduce scientific and Army resource demands. The following activities will be carried out by the Task 5 data collection team for other project tasks during cohort data collection:

(1) Administer, in line with guidance and training given by Task 2 staff, the four-hour predictor battery (second cohort test administration only).

(2) Administer the Task 3 job-knowledge tests for all 19 MOS (first cohort test administration only).

(3) Collect, under the supervision of Task 4 staff, Army-wide measures and ratings (all cohort administrations), as well as utility judgments for MOS performance levels.

<u>Activity 5.7.5 Analyze main cohort data and report results</u>. We will forward the MOS-specific criterion performance data to the LRDB for processing and analysis. Task 1 staff will be responsible for carrying out the array of validity analyses that are planned. Task 5 staff will also conduct a variety of analyses using the cohort data, but these analyses will be developmental in nature and designed to improve the quality of the performance tests.

Many of the analyses described for the field test data (Activity 5.5.5) will be conducted again on the largest cohort samples. The principal purposes of these analyses will be to reestablish that the hands-on and job

knowledge performance tests as well as the behavioral and overall rating scales are sound, useful, efficient and cost-effective criterion measures. We will reestimate reliabilities, compute estimates of concurrent and construct validity for job knowledge tests and rating scales, make certain that all of the criterion instruments are in compliance with APA Division 14 guidelines, and factor analyze the correlations among instruments in a continuing effort to refine the criterion battery. If the analyses identify any aspects of the measures that warrant improvement, the necessary modifications will be made after receiving the concurrence of the COR.

Report results. We will conclude the test of each MOS group in the main cohort with a report that sets forth the methods, samples, variables, instruments, and results of the test. The report will present the component set of measures tested and the results of the analyses (e.g., psychometric properties, utility). Especially for MOS groups tested early in the project, the report will discuss the implications for tests in subsequent MOS groups (e.g., use of automated devices for data collection, tasks common to more than one MOS).

The reports will contain management summaries for the portion of the audience that needs to make decisions based on the validities, utilities, and other management information. The body of the report will meet the criteria for publication by ARI or by psychological journals, including details of research design, theoretical bases, results and discussions.

Support requirements for Subtask 5.7. In the cohort data collection we are concerned with the effect of sample size on two major statistical

issues in addition to those mentioned for the field test. Chief among these issues is the question of differences in performance or validity for subgroups of particular interest (women, Blacks, and Hispanics). With an overall sample size of 650 soldiers in each MOS, we expect to sample where appropriate at least 120 soldiers from each of the key subgroups. With subgroup samples of this size, observed subgroup differences of 10 percent in the proportion passing individual items or tasks would be statistically significant. Since differences of 10 percent in the proportion passing individual items are of practical significance, it is essential that the samples be large enough to detect these differences reliably. In addition, more accurate estimates of percent passing and correlations among measures are required for cohort measurement than for field testing the measures. The proposed sample size will give us 95 percent confidence bounds of plus or minus 4 percentage points in estimates of percent passing, and 95 percent confidence bounds of plus or minus .08 for estimates of correlations based on the entire sample. While more accurate estimates would clearly be desirable, this level of accuracy is judged acceptable for the purposes of the project.

Soldier support requirements for implementing data collection on SL1 MOS-specific performance measures in the FY83/84 cohort are shown in Table 5.7.1. Soldier support requirements for implementing data collection on MOS-specific performance measures in the FY86/87 cohort and SL2 measures in the FY83/84 cohort are shown in Table 5.7.2. Soldier support requirements for implementing data collection on SL2 MOS-specific performance measures in the FY86/87 cohort are shown in Table 5.7.3.

Table 5.7.1

Soldier Support Requirements for Cohort Test I
(Task 5 measures only)

Purpose	Soldiers MOS	Soldiers SL	Number	Total	Days Per Person
Provide liaison with tested units	Officer		1 per site	16	30
Coordinate equipment and subjects	NA	3-5	3 per site	48	10
Score hands-on tests	9 MOS	2	18 per site [2 per MOS]	288	30
Subjects for performance measures	9 MOS	1	365 per site [42 per MOS]	6000	1
Supervisor ratings	Supervisor	3-5	100 per site	1600	.5

Table 5.7.3

Soldier Support Requirements for Cohort Test II
(Task 5 measures only)

Purpose	Soldiers MOS	Soldiers SL	Number	Total	Days Per Person
Provide liaison with tested units	Officer		1 per site	16	30
Coordinate equipment and subjects	NA	3-5	3 per site	48	10
Score hands-on tests	9 MOS	3-5	18 per site [1 per MOS]	288	30
SL1 subjects performance measures	9 MOS	1	315 per site [35 per MOS]	5000	1
SL2 subjects for performance measures	9 MOS	2	56 per site [6 per MOS]	900	1
Supervisor ratings	Supervisor of subjects	3-5	100 per site	1600	.5

Table 5.7.3

Soldier Support Requirements for Cohort Test III

Purpose	Soldiers MOS	Soldiers SL	Number	Total	Days Per Person
Provide liaison with tested units	Officer		1 per site	16	10
Coordinate equipment and subjects	NA	3-5	3 per site	48	3
Score hands-on tests	9 MOS	3-5	9 per site	144	7
Subjects for performance measures	9 MOS	2	56 per site [6 per MOS]	900	1
Supervisor ratings	Supervisor of subjects	3-5	18 per site	300	.5

Subtask 5.8: Produce Final Technical Report on MOS-Specific Performance Measures

Reports on MOS-specific performance measures include the reports on the field and cohort tests for each MOS group, a report describing each of the instruments and measures, a final technical report, and articles for publication.

We will prepare the draft final technical report in two parts. The executive summary will present an overview of the purpose, procedures, results and use of the performance measures in a brief text with graphic displays suitable for managers and Army operational personnel. An expanded scientific section will present details of methodology, development of the performance measures, technical problems and their resolution, statistical analyses, results, interpretations, and use.

This final report will synthesize the research for all of the MOS groups. It will discuss the value and utility of the two classes of measures in the project: "ultimate" measures of task performance and "proximal," indirect, low-cost measures (e.g., pre-induction predictors, school/training measures). The indirect class of measures will be described as to their empirical relationships with the "ultimate" measures and other variables in the research (e.g., predictors, Army-wide measures).

Task 5 activities per se, including an instrument/measurement report and a final report, are scheduled to be completed at the end of the 7th project year. Project A staff will collect the data and conduct the analysis for the final cohort sample (FY86/87 cohort, second tour). The findings will be incorporated into an addendum to the final report.

SUMMARY OF EXPECTED OUTCOMES

The work of Task 5 is expected to produce a number of outcomes that will have value beyond the basic purpose of the task. Some of these outcomes will be of scientific interest, applicable to other researchers and other situations; others will be of practical value, primarily in the military services but to other applications as well. Most of the outcomes, however, will have both scientific and operational value, since the scientific outcomes will, for the most part, be developed to solve an operational problem. In the paragraphs that follow, we have separated the outcomes into scientific and operational categories, though even the operational outcomes would be of interest to other researchers faced with similar problems.

Scientific Outcomes

1. A primary scientific contribution will be the taxonomy of behaviors and constructs that span the range of Army MOS, and the techniques developed to produce the taxonomy. For the first time, such a taxonomy of human performance will be built on large samples of data from a wide array of jobs performed by a sample of men and women relatively homogeneous in age, training, and experience, but who differ in aptitudes and adaptation to work and Army life. Future Army research will not have to rely on meta-analyses of old data collected for different purposes, or on small heterogeneous samples collected under various sets of instructions or conditions.

Tasks 5 and 2 will be able to identify the similarities and dissimilarities in the job activities and the KSAO required to perform those activities. The longitudinal aspect of the research will reflect changes in the activities and the KSAO as the soldier moves through the enlisted ranks. Such a taxonomy will be of benefit for years to come -- to both the military and civilian workforce, to long-range manpower planners, and to researchers.

2. A procedure will be developed for selecting tasks for testing that addresses many of the questions that have plagued job performance measurement. The procedure will be a practical approach to selecting tasks for large-scale job proficiency assessment efforts in both public and private sectors. The military services may need to modify the approach to achieve training benefits, but the procedure developed by this project is likely to provide a foundation for future measurement programs.

One of the goals of our research is to develop a procedure to select tasks and measures for the specific purpose of "generalization to other tasks and MOS." In order to meet this objective, we will use various techniques, ratings, categorizations, and taxonomies to analyze a large number of MOS and tasks. While we will make use of these analyses for the specific purposes of the project, there are many other potential uses for these results.

For example, suppose we ultimately decide to develop front-end task analyses for large numbers of MOS and/or tasks to be used to select MOS/task

for further research or generalization to Army-wide operations. Obviously, such analyses can be used by appropriate audiences in MOS other than the ones we select to investigate in detail. More generally, our task analysis procedures will be made "public" in the sense that they will be user-oriented, definitions will be carefully operationalized, and procedures will be standardized so that they could be used wherever desired in the Army.

In addition, the specific decision rules we adopt for our purposes -- i.e., task selection for representativeness and generalization -- will have direct application wherever and whenever similar decisions are made. For example, two ubiquitous problems for Army trainers are how to select tasks for inclusion in training, and how to select tasks for readiness evaluation. At the core of these problems is the identical issue of representativeness and generalization. A standardized, operational codification of whatever procedures and rules we develop would be of direct value.

3. An empirically based definition will be developed of the job performance space in which skilled and motivational aspects of job specific performance are articulated with respect to training achievement and other more general (e.g., "Army-wide") indicators of performance.

Traditionally, job/task analysis has tended to focus on relatively discreet relationships between training performance and subsequent job performance, between individual aptitude and training, or between aptitude and performance. Similarly, "performance" has tended to be viewed either as

proficiency in certain skills or, alternatively, as the achievement of more global goals related to mission accomplishment or other organizational outcomes. One of the major outcomes of Project A should be the integration of these elements into a reasonably coherent model.

This is also a matter of some interest from an operational standpoint. In large organizations such as the Army, where initial training is largely separated from day-to-day operations in the field, there is a strong tendency for the two functions to become less and less related to each other. This problem manifests itself most visibly in the tendency of operational personnel to discount or dismiss what is taught in training as irrelevant or inadequate. At least part of the reason for this divergence is the lack of a common language to describe the job. Project A presents an opportunity to begin developing such a vocabulary within several MOS, language that can be extended not only to other Army jobs, but also to similar civilian jobs.

4. Evidence will be acquired as to the relative efficiency of alternative methods of job performance measurement. Since some aspects of job-specific performance will be measured by more than one method, and since method development and administration costs will be recorded, cost-benefit can be analyzed.

By the end of this project, we will know a great deal about the possibilities and limitations of a broad variety of job performance techniques. And, because of the variety of jobs and job settings we will encounter, we will have the additional perspective of having attempted to

develop measures for many types of skills and attributes, including technical, mechanical, cognitive and physical requirements. In short, we should be in a position to write the "handbook" on job performance measurement. Such a document should be able to compare techniques, costs, problems of measurement development and implementation, test validity and reliability, considerations for special populations, and the like. Data will be obtained concerning the suitability of measurement methods/media to people with different aptitudes. For example, is a written job knowledge test as valid a job performance measure for those low in verbal aptitude as it is for those high in verbal aptitude?

Project A will generate a validated procedure for developing indirect but feasible proficiency measures, such as performance ratings or knowledge tests. The project offers a unique opportunity to examine the practical question of the reliability of indirect measures of job performance, (e.g., peer ratings or paper-and-pencil tests) as compared to more expensive direct measurement procedures (e.g., hands-on testing). In addition, the project will examine these relationships for a large number of tasks, reflecting many different types of skills and knowledge areas. We should be able to state the conditions under which less-direct measurement methods can be applied, the level of reliability to be expected, how such measures can and should be developed and the acceptable limit of such methods. From the Army's standpoint, this knowledge could be applied to evaluate existing efficiency rating systems or to develop new, more reliable systems. Moreover, the project offers a chance to relate these individual assessments to unit- or higher-level performance.

An algorithm will be developed for matching test method to test characteristics. Such an algorithm will allow test developers with subject matter expertise to take advantage of the results of our cost-benefit analyses without undertaking the field trials. The algorithm will also enhance the efficiency of future large-scale job proficiency measurement programs.

5. A better understanding of the relationship between school-related skills and knowledges and job-related skills and knowledges should emerge from this work. If the predictors used relate to one and not the other, a reassessment of the school curriculum or the criterion performance measures might be suggested. Since we will be developing our own performance tests, a situation could arise that would point to the school as the source of a problem. The table below shows the possible interactions between the school and job performance. Cells A and D are the two that suggest a school problem.

	PERFORMANCE ON THE JOB	
School Performance	Low	High
High	A	B
Low	C	D

The role of factors not related to job performance per se, but to the contextual factors in which the job is carried out, should be clarified. To what extent does poor job performance (as in cells A and C above) relate to these factors? How can these factors be measured for criterion purposes? Cell D suggests that some people, at least, can perform well despite poor school performance; what contextual factors account for the

improvement? Clearly, major changes in curricula and other factors may be needed to substantially raise the percentage of soldiers who do well on the job.

An important potential for the research is the opportunity to explore the relationship among the knowledges, skills, and abilities to do the job ("can do") and the soft skills, or "will do" aspects of the world of work. While the project may not find the solution, it should produce correlates of the soft skills that could be used in further studies aimed at this elusive goal.

6. A better understanding of test bias for race and gender should result from the project. Since the MOS have been selected to include substantial representation by Blacks, Hispanics, and women, field data will provide a unique opportunity to study the interplay of task type, test method and job experience for the various subgroups.

Operational Outcomes

1. The MOS taxonomy will be defined in terms of similar job activities, which may be used to augment or refine Career Management Fields as a personnel management tool. Current MOS descriptions suffer from a lack of standardized terms, definitions and conceptual categories. Tasks, duties and duty positions are described differently between and within MOS. The MOS taxonomy will provide a consistent, unified and standardized system and language for describing jobs, tasks and duties, as well as the underlying skills and knowledge categories needed to perform those tasks. This will

make it possible to compare and possibly refine MOS. Such an outcome is a prerequisite for expanding the findings of this project beyond the 19 MOS studied in Project A.

2. An appraisal will be made of existing job/task analytic systems employed by the Army, such as the AOSP survey. The project provides an opportunity to view the curre ASOP process across several MOS. Several by-products of this review would include: a comparison of ASOP job descriptions with other authoritative descriptions (e.g., AR 611-201, Soldier's Manuals); the identification of redundancies and varying levels of detail within and between job descriptions; improved consistency in the use of terms and concepts among several surveys. Beyond this, the project could provide a basis for standardizing some aspects of the surveys such as the polling of incumbents for specific skills, knowledge and physical requirements of their jobs.

3. The knowledge of the relationship between ASVAB measures -- new and old -- and performance on tasks will also be of great value to the military in developing test profiles for MOS not included in the study and for emerging MOS. Since the validity of the ASVAB tests for a wide range of tasks will be known, additional or emerging MOS can be identified in terms of those tasks, and an appropriate ASVAB profile determined. For example, suppose MOS were described by tasks as follows:

Task	MOS											
	A	B	C	D	E	F	G	H	.	.	.	n
1	x	x	x		x		x					
2	x	x	x				x	x				
3	x		x		x		x	x				
4	x	x			x		x					
5		x	x	x	x			x				
6		x		x	x		x	x				
7		x		x				x				
8						x						
9						x	x					
.												
.												
.												
n												

Tests valid for MOS-A should also be valid for MOS-G, but in addition, the test associated with Tasks 6 and 9 should also be used for MOS-G. If such a modular approach can be developed, because of known validities of tests for tasks, the testing for MOS in the future can become more directed, and eventually computerized. As an enlistment applicant passes a test, all MOS that include the task(s) covered by that test are options for him/her. Upon completion of the testing, the recruiter will have an applicant profile and a printout of all MOS for which the applicant is eligible, and the "best fits" in terms of test scores and/or expected net utilities.

4. The measurement instruments themselves, produced in the project will be an important outcome. Though developed for only a sample of MOS, the tests and rating instruments could serve as measures of training achievement and job proficiency.

For the particular MOS investigated, Task 5 research will result in a large set of reliable, valid, and user-tested instruments, designed in a scientifically sound manner. Thus, instrument administration procedures,

task conditions and standards of performance, equipment requirements, scoring rules, and scorer qualifications and procedures will have been codified and standardized. They will be immediately useful for evaluating training and job proficiency, diagnosing weaknesses and strengths in school training, diagnosing requirements for OJT or refresher training, and unit readiness. The instruments can have other uses in units. For example, unit Training Officers (or unit Commanders) might use a test battery to screen arrivals for specific duty assignments, or use such a battery to plan unit training exercises, or to structure individual soldier advancement criteria (in lieu of SQT).

5. The procedures used to produce a test or other measurement instrument will be made public and can be used by appropriate people in other MOS. There is nothing mystical or esoteric about producing a test; the steps involved (e.g., listing individual steps, deciding on conditions and standards, trying-out procedures, reliability checks, etc.) could easily be performed by school and/or unit personnel. We could produce a "how-to" manual, usable by anyone who desires a good test. This manual would be available to all appropriate audiences.

Other unforeseen by-products of scientific or operational interest are likely to result from the Task 5 research. Some of these may be born of serendipity, others of need. The former, by their nature, cannot now be described. The latter, on the other hand, are illustrated by our reanalysis of the Army occupational domain, which was undertaken out of a need to classify MOS for sampling purposes--work that may shed some light on the general skill requirements underlying the Army's Career Managment Fields.

REFERENCES

Angell, D., Shearer, J.W., & Berliner, D.C. Study of training performance evaluation techniques (TR-NAVTRADEVCEN-1449-1). Orlando, FL: Naval Training Device Center, 1964.

Angoff, W.H. Test reliability and effective test length. Psychometrika, 1953, 28, 1-14.

Borman, W.C. Individual differences correlates of accuracy in evaluating performance effectiveness. Applied Psychological Measurement, 1979, 3, 103-115.

Borman, W.C., Dunnette, M.D., & Johnson, P.D. The development and evaluation of a behavior-based naval officer performance assessment package. Minneapolis: Personnel Decisions, Inc., 1974.

Borman, W.C., Hough L.M., & Dunnette, M.D. Development of behaviorally based rating scales for evaluating the performance of U.S. Navy recruiters (TR-76-31). San Diego, CA: Navy Personnel Research and Development Center, 1976.

Brumback, G.B., Romashko, T., Hahn, C.P., & Fleishman, E.A. Model procedures for job analysis, test development and validation. Washington, D.C.: American Institutes for Research, 1974.

Campbell, J.P., Dunnette, M.D., Arvey, R.D., & Hellervik, L.V. The development and evaluation of behaviorally based rating scales. Journal of Applied Psychology, 1973, 57, 15-22.

Christal, R.E. (Chm.) Collecting, analyzing and reporting information describing jobs and occupations. Symposium presented at the meeting of the American Psychological Association, Washington, D.C., September 1969.

Christal, R.E. The United States Air Force occupational research project (AFHRL-TR-73-75). Brooks Air Force Base, Texas: Air Force Human Resources Laboratory, January 1974.

Cronbach, L.J., Gleser, G.C., Nada, H., & Rajartman, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Department of the Army. Training management in battalions (TC 21-5-7). Washington, D.C.: Author, 1977.

DeMaio, J., Parkinson, S., Leshowitz, B., Crosby, J., & Thorpe, J.A. Visual scanning: Comparisons between student and instructor pilots. AFHRL-TR-76-10, AD-A023634. Williams AFB, AZ: Flying Training Division, Air Force Human Resources Laboratory, June 1976.

Engel, J.D. An approach for standardizing human performance assessment. Paper presented at THEMIS Conference, Lubbock, Texas. October 1970.

Engel, J.D., & Rehder, R.J. A comparison of correlated job and work sample measures for general vehicle repairmen (HumRRO Technical Report 70-16). Alexandria, Virginia: Human Resources Research Office, October 1970.

Fine, S.W. Functional job analysis. Personnel Administration and Industrial Relations, 1955, 2, 1-16.

Flanagan, J.C. The critical incident technique. Psychological Bulletin, 1954, 51, 327-358.

Fleishman, E.A. On the relation between abilities, learning, and human performance. American Psychologist, 1972, 27, 1017-1032.

Foley, J.P., Jr. Evaluating maintenance performance: An analysis (AFHRL TR-64-57(I)). Wright-Patterson Air Force Base, Ohio: Air Force Human Resources Library, 1974.

Frederickson, U. Proficiency tests for training evaluation. In R. Glaser (Ed.), Training research in education. Pittsburgh, Pennsylvania: University of Pittsburgh Press, 1962.

Harris, A., & Mackie, R.R. Factors influencing the use of practical performance tests in the Navy. ONR Technical Report 703-1, August 1962.

Harris, J.H., Campbell, R.C., Osborn, W.C., & Boldovici, J.A. Development of a model job performance test for a combat occupational specialty: Volume I. Test development; Volume II. Instructions and procedures for conducting a functionally integrated performance test (HumRRO Final Report FR-CD(L)-75-6). Alexandria, Virginia: Human Resources Research Organization, November 1975.

Joreskog, K.G., & Sorbom, D. LISREL V: Analysis of linear structural relationships by the method of maximum likelihood (Research Report 81-8. Uppsala, Sweden: University of Uppsala, 1981.

Klein, G.A. Phenomenological approach to training. AFHRL-TR-77-42. Wright-Patterson Air Force Base, OH: Advanced Systems Division. August 1977.

Klein, G.A. Problems and opportunities in deriving training requirements for design and utilization of simulators. Proceedings of the First International Learning Technology Congress and Exposition, July 1976.

Knoop, P.A., & Welde, W.L. Automated pilot performance assessment in the T-37: A feasibility study. AFHRL-TR-72-6, Ad-766 446. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, April 1973.

Madden, J.M., Hazel, J.T., & Christal, R.E. Worker and supervisor agreement concerning the worker's job description. Personnel Research Laboratory Aerospace Medical Division, Air Force Systems Command, 1964.

March, J.E., & Archer, W.B. Procedural guide for conducting occupational surveys in the United States Air Force (Technical Report No. 67-11). Lackland Air Force Base, Texas: Personnel Research Laboratory, September 1967.

McCormick, E.J., Jeanneret, P.R. & Mecham, R.C. A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). Journal of Applied Psychology, 1972, 56, 347-368.

Miller, R.B. Task analysis: Sources and futures. Improving Human Performance, 1973, 2, 5-27.

Miller, R.B. A method for man-machine task analysis (WADC Technical Report 53-137). Wright-Patterson Air Force Base, Ohio, June 1953.

Montemerlo, M.D., & Eddowes, E. The judgmental nature of task analysis. Proceedings of the Human Factors Society, 22nd Annual Meeting, 1973.

Osborn, W.C., & Ford, J.P. Research on methods of synthetic performance testing (HumRRO, Final Report FR-CD(L)-76-1). Alexandria, Virginia: Human Resources Research Organization, January 1977.

Prelewicz, S.J. A report on the state of the art of task analysis. Unpublished paper, April 5, 1977.

Rose, A.M., Shettel, H.H., Wheaton, G.R., Bolin, S.F., & Barba, M.A. Evaluating the effectiveness of soldier's manuals: A Field Study. Washington, DC: American Institutes for Research, February 1981.

Rupe, J.C. Research into basic methods and techniques of Air Force job analysis - IV (AFPTRC-TN-56-51). Lackland Air Force Base, Texas: Air Force Personnel and Training Research Center, April 1956.

Schmidt, F.L., Greenthal, A.L., Hunter, J.E., Berner, J.G., & Seaton, F.W. Job sample versus paper and pencil trades and technical tests: Adverse impact and examinee attitudes. Personnel Psychology, 1977, 30 (2), 187-197.

Shirkey, E.C. Preliminary validity report of the MOS evaluation test for medical specialist (USAEEC Research Study No. 31). Ft. Benjamin Harrison, Indiana: U.S. Army Enlisted Evaluation Center, 1965.

Shirkey, E.C. Preliminary validity report of the MOS evaluation test for general supply specialist (USAEEC Research Study No. 38). Ft. Benjamin Harrison, Indiana: U.S. Army Enlisted Evaluation Center, 1966.

Toquam, J.L., & Borman, W.C. Development of first line supervisor behavior summary scales. Minneapolis, Minnesota: Personnel Decisions Research Institute, 1981.

Urrey, V.W., Shirkey, E.C., & Nicewander, W.O. A preliminary validity report of the MOS evaluation test for personnel specialist, MOS Code 716.1 (USAEEC Research Study No. 21). Ft. Benjamin Harrison, Indiana: U.S. Army Enlisted Evaluation Center, 1965.

Urrey, V.W., Shirkey, E.C., & Waldkoetter, R. O. Preliminary validity reporting of the MOS evaluation test for cook (USAEEC Research Study No. 20). Ft. Benjamin Harrison, Indiana: U.S. Army Enlisted Evaluation Center, 1965.

U.S. Army Continental Army Command. Training: Systems engineering of training (course design). CON Reg 350-100-1, 1972.

U.S. Army Continental Army Command. Training: Systems engineering of training (ATP, ATT, Non-MOS ASUBJSCD). CON Pam 350-11, 1973.

U.S. Army Training and Doctrine Command. Collective front-end analysis (CFEA) for development of Army training and evaluation program (ARTEP). TRADOC Pam 310-8 (Draft), undated.

U.S. Army Training and Doctrine Command. U.S. Army interservice procedures for instructional systems development. TRADOC Pam 350-30, 1975.

U.S. Army Training and Doctrine Command. Job and task analysis handbook. TRADOC Pam 351-4(T), 1979.

Van Cott, H.P., Berkun, M.M., & Purifoy, G.R., Jr. How to determine job knowledge content: The application of a procedure to several Air Force jobs. Pittsburgh, Pennsylvania: American Institutes for Research, December 1955.

Vineberg, R., & Taylor, E. Performance in four Army jobs by men at different aptitude (AFQT) levels: 4. Relationships between performance criteria (HumRRO Technical Report 72-73). Alexandria, Virginia: Human Resources Research Office, August 1972.

'ellen, T. Preliminary validity report of the MOS evaluation test for track vehicle mechanic, MOS Code 632.2 (USAEEC Research Study No. 49). Fort Benjamin Harrison, Indiana: U.S. Army Enlisted Evaluation Center, 1966.